NAVAL POSTGRADUATE SCHOOL

MONTEREY, CALIFORNIA

THESIS

DEVELOPING A SCALED PERFORMANCE EVALUATION MEASUREMENT SYSTEM TO EVALUATE MARINE PERFORMANCE

by

Garrett A. Loeffelman

June 2019

Thesis Advisor: Glenn A. Hodges
Co-Advisor: Quinn Kennedy

Research for this thesis was performed at the MOVES Institute.

THIS PAGE INTENTIONALLY LEFT BLANK

REPORT DOCUMENTATION PAGE			Form Approved OMB No. 0704-0188

1. AGENCY USE ONLY *(Leave blank)*	2. REPORT DATE June 2019	3. REPORT TYPE AND DATES COVERED Master's thesis

4. TITLE AND SUBTITLE DEVELOPING A SCALED PERFORMANCE EVALUATION MEASUREMENT SYSTEM TO EVALUATE MARINE PERFORMANCE	5. FUNDING NUMBERS RVM1M
6. AUTHOR(S) Garrett A. Loeffelman	
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) Naval Postgraduate School Monterey, CA 93943-5000	8. PERFORMING ORGANIZATION REPORT NUMBER
9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES) N/A	10. SPONSORING / MONITORING AGENCY REPORT NUMBER

11. SUPPLEMENTARY NOTES The views expressed in this thesis are those of the author and do not reflect the official policy or position of the Department of Defense or the U.S. Government.

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

13. ABSTRACT (maximum 200 words)

Training developers lack methods for determining the benefits of integrating live, virtual, and constructive training. This study defined and tested a scaled performance evaluation measurement system (SPEMS) to be used across tasks. We used the buddy rush task to test SPEMS and compare it to the current "Go/No Go" performance evaluation checklist (PECL). We developed SPEMS in three steps: we convened focus groups to establish five-level behaviorally anchored rating scales (BARS); confirmed SPEMS reliability using subject-matter expert (SME) virtual video analysis; and empirically tested SPEMS' predictive capability in an operational environment. Suitable inter-rater reliability was found for BARS (87% agreement) and SPEMS (Cronbach's Alpha 0.93 to 0.98). Percent exposure was selected by SMEs as the objective measure of buddy rush performance. Fifty-two trainees (26 pairs) were evaluated using a PECL and SPEMS at three time points. The results showed that SPEMS has a moderate, negative, linear relationship with percent exposure at an R2 = 0.41/0.40. Conversely, PECL has a weak, slightly negative linear relationship with percent exposure at an R2 = 0.03/0.2. We reject the null hypotheses and conclude that SPEMS scores are significantly related to percent exposure and have more predictive strength than PECL scores. These findings demonstrate a verifiable, repeatable, and reliable potential solution to the problem of measuring military task performance across training solutions.

14. SUBJECT TERMS evaluation, human performance, LVC, metrics, operational environment, proficiency, readiness, return on investment, training	15. NUMBER OF PAGES 133
	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.

DEVELOPING A SCALED PERFORMANCE EVALUATION MEASUREMENT SYSTEM TO EVALUATE MARINE PERFORMANCE

Garrett A. Loeffelman
Captain, United States Marine Corps
BS, U.S. Naval Academy, 2013

Submitted in partial fulfillment of the
requirements for the degree of

MASTER OF SCIENCE IN MODELING, VIRTUAL ENVIRONMENTS, AND SIMULATION

from the

NAVAL POSTGRADUATE SCHOOL
June 2019

Approved by: Glenn A. Hodges
Advisor

Quinn Kennedy
Co-Advisor

Peter J. Denning
Chair, Department of Computer Science

THIS PAGE INTENTIONALLY LEFT BLANK

ABSTRACT

Training developers lack methods for determining the benefits of integrating live, virtual, and constructive training. This study defined and tested a scaled performance evaluation measurement system (SPEMS) to be used across tasks. We used the buddy rush task to test SPEMS and compare it to the current "Go/No Go" performance evaluation checklist (PECL). We developed SPEMS in three steps: we convened focus groups to establish five-level behaviorally anchored rating scales (BARS); confirmed SPEMS reliability using subject-matter expert (SME) virtual video analysis; and empirically tested SPEMS' predictive capability in an operational environment. Suitable inter-rater reliability was found for BARS (87% agreement) and SPEMS (Cronbach's Alpha 0.93 to 0.98). Percent exposure was selected by SMEs as the objective measure of buddy rush performance. Fifty-two trainees (26 pairs) were evaluated using a PECL and SPEMS at three time points. The results showed that SPEMS has a moderate, negative, linear relationship with percent exposure at an R2 = 0.41/0.40. Conversely, PECL has a weak, slightly negative linear relationship with percent exposure at an R2 = 0.03/0.2. We reject the null hypotheses and conclude that SPEMS scores are significantly related to percent exposure and have more predictive strength than PECL scores. These findings demonstrate a verifiable, repeatable, and reliable potential solution to the problem of measuring military task performance across training solutions.

THIS PAGE INTENTIONALLY LEFT BLANK

TABLE OF CONTENTS

THIS PAGE INTENTIONALLY LEFT BLANK

LIST OF FIGURES

THIS PAGE INTENTIONALLY LEFT BLANK

LIST OF TABLES

THIS PAGE INTENTIONALLY LEFT BLANK

LIST OF ACRONYMS AND ABBREVIATIONS

03xx	Infantry Marine
AAR	After action review
ADDIE	Analyze, design, develop, implement, and evaluate
AGTS	M1A1 Advanced Gunnery Training System
ANOVA	Analysis of variances
BARS	Behaviorally anchored rating scale
CRP	Computed readiness percentage
DoD	Department of Defense
G-Theory	Generalizability Theory
HBCT	Heavy brigade combat team
HRPP	Human research performance program
IO	Instructor operator
IRB	Institutional review board
ISD	Instructional system design
ISTS	Integrated Simulation Training System
ITEAM	Integrated training environment assessment methodology
ITS	Individual training standard
JPM	Joint-Service Job Performance Measurement project
LF	Live Fire
LVC	Live, virtual, constructive
MET	Mission essential task
METL	Mission essential task list
MOE	Measure of effectiveness
MOP	Measure of performance
MOS	Military occupational specialty
MOVES	Modeling of virtual environments and simulation
OJT	On the job training
PECL	Performance evaluation checklist
PME	Professional military education
ROI	Return on investment
SAT	*The Systems Approach to Training*
SME	Subject matter expert
SOI	School of Infantry

SPEMS	Scale performance evaluation measurement system
START	Systematic team assessment of readiness training
T&R	Training and Readiness
TEE	Training effectiveness evaluation
UTM	*Unit Training Management Guide*
VBS3	Virtual Battle Space 3

ACKNOWLEDGMENTS

I would like to thank a number of people who have assisted and guided me through the development of this thesis work. First, to my ever-supportive wife, Sarah, thank you for always listening to me, challenging me, and being by my side. Your commitment to our family is ironclad, and I am truly blessed to have such an intelligent and beautiful confidant. I promise you won't have to hear any more about "My Thesis." To my first thesis advisor, Dr. Glenn Hodges, LTC, USA, I appreciate your commitment to excellence and for forcing me to take the more difficult road. You are one of the most academically driven military officers I have ever met, and I am deeply grateful that you impressed upon me the importance of scientific and literary rigor. To my second thesis advisor, Dr. Quinn Kennedy, you are a consummate professional, the most brilliant professor I have ever met, and the smartest person in any room. You have forever changed the way I will look at the world, and I consider myself honored to have worked with you on this thesis. To my mom, Pam Loeffelman, you're my hero. Your dedication to professionalism, hard work, and personal development are inspiring. To my dad, Bob Loeffelman, thank you for teaching me what it means to be a gentleman. You raised me, and provided the balance, sense of self, and confidence that is required to succeed in this ever-changing world. To my brother, Brenton, thank you for always being my academic superior, for challenging me, and for always paving the way ahead. I would be remiss if I did not mention my mother- and father-in-law who have always welcomed me into their family as a son; thank you for your unending support. To LtCol Day and the training staff at the School of Infantry–West at Camp Pendleton, CA, thank you for enabling this important work. I would like to extend a personal thank you to the Naval Postgraduate School support staff who made this work a reality: Rabia Khan, Eric Johnson, and Ryan Lee. Your effort throughout this process was unwavering and critical to the success of this work. Finally, to my friends and fellow cohort members, I will miss you all; every one of you has matured a different part of me that has helped mold and shape my endeavors indelibly. Last, but certainly not least, I must thank my dog, Norman. He is my most loyal friend, who has sat by my feet through this entire process. Thanks buddy.

THIS PAGE INTENTIONALLY LEFT BLANK

I. INTRODUCTION

A. PURPOSE

The Marine Corps is interested in a consistent system that measures how Marines perform tasks to quantitatively demonstrate the benefits of implementing new training programs. Today, collective and individual tasks are measured using binary performance evaluation checklists that inform Marines if they are trained in a task. However, these measures do not inform Marines of their level of proficiency. A measurement system that is capable of not only comparing Marines to their peers, but also a training system's ability to improve a Marine's performance of tasks is desired. Such a system might use a scaled performance evaluation measurement system (SPEMS) to show the benefits of training in different environments. This thesis offers such a methodology.

This thesis has three goals. The first is to define and develop SPEMS for classifying an individual's performance given a specific task. The second is to test the value and usability of SPEMS. For this, we designed and executed an experiment that assesses whether SPEMS accurately predicts and captures an individual's performance on a given task. The third and exploratory goal is to provide a method for calculating the return on investment of a simulator using the performance quantified by SPEMS.

The development of SPEMS has the potential to measure task performance proficiency on a graduated scale instead of the current binary "Go/No go" evaluation technique. SPEMS could allow training developers to compare specific systems more objectively. Quantitatively comparing Marine task performance across different training systems could enable developers to determine which system provides the optimal solution. These stated goals are addressed in this document sequentially over the following six chapters.

B. CHAPTER OUTLINE

Chapter II addresses the main problems associated with quantifying military task performance within the current environment. This chapter starts by studying the specific

task that was tested during experimentation, the buddy rush task. The chapter concludes with an in-depth examination of how tasks like the buddy rush task are typically trained and evaluated to demonstrate where improvements can be made. Outlining the problem enables us to move into the background section with an understanding of what challenges need to be addressed.

Chapter III focuses on background information related to solving the problems outlined in Chapter II. Chapter III provides the history behind military tasks and highlights what successful military performance evaluation should look like. The chapter dispels bias and reliability concerns, as well as introduces some performance evaluation methods. The chapter concludes with a proof of concept study that demonstrates the relevance of creating a performance evaluation method within the context of the problem. This chapter serves as a literature road map that explains how this study was conducted.

Chapter IV addresses the first goal of this thesis by outlining the pilot study methodology that researchers developed and implemented. This chapter begins with a thorough explanation of the pilot study and ends with a fully developed SPEMS. Once SPEMS was fully developed, we conducted live-experimentation to validate SPEMS.

Chapter V addresses the second goal of this thesis by describing the live experiment that was conducted in the operational environment to validate SPEMS. This chapter details the methodology and concludes with the experiment's results. These results serve as a solution to the challenges laid out in Chapters I and II and a basis for defining our conclusions.

Chapter VI addresses the third goal by detailing the discussions, recommendations and future work sections of the thesis. This chapter focuses on analyzing the results within the context of the problem and concludes with our view of how this research demonstrates a verifiable, repeatable, and reliable potential solution to the problem of measuring military task performance.

II. PROBLEM STATEMENT

A. INTRODUCTION

The USMC lacks adequate performance data to determine the benefits of integrating live, virtual, and constructive (LVC) simulation capabilities into our current training programs. Additionally, training developers lack appropriate methods for determining the benefits of integrating simulation-based training solutions. These issues were highlighted in the Government Accountability Office Report 13–698 on Army and Marine Corps Training, *Better Performance and Cost Data Needed to More Fully Assess Simulation-Based Efforts*. The primary reason the USMC lacks adequate performance data is that the performance of military tasks is challenging to quantitatively measure. We examined one specific task as an example to highlight why it is difficult to measure performance and evaluate the training of complex tasks.

B. THE INDIVIDUAL TRAINING STANDARD—A WEAK PERFORMANCE EVALUATION USE CASE

The individual training standard (ITS) defined by the USMC, provides the framework for how the USMC evaluates performance. ITSs are a part of a commander's mission essential task list (METL). METL development is the process where a commander chooses what tasks his unit must be capable of proficiently accomplishing. The training and readiness (T&R) manuals establish a hierarchical approach to tasks to allow commanders to choose specific unit level tasks (individual through regiment) (Marine Corps, 2011). Each higher-level T&R task is chained to the subordinate tasks responsible for making up the mission essential task. A couple of core competencies at the battalion and regimental levels breaks down to thousands of individual, fire team, squad, platoon, company, and battalion tasks (Marine Corps, 2011). ITSs are the backbone of every core competency to which the Marine Corps is required to train, and they provide the framework for how a unit's readiness is evaluated. Unit performance is measured by aggregating the inaccurate binary "Go/No Go" task performance evaluation into a computed readiness percentage (CRP). Tasks are evaluated as a "Go" if evaluators determine the tasks were conducted to standard. Wong et al. noted that averaging these varied performance standards

typically confounds the accuracy of the computed unit readiness score (Wong, Gerras, & Barracks, 2015). The binary system only allows commanders to succeed or fail without providing a granular feedback system. This system fosters an environment where evaluators are pressured to report that all tasks evaluated were trained to standard regardless of how well they were actually performed (Wong et al., 2015). *Due to the fact that all commander's METLs are made up of T&R tasks, a dysfunctional task evaluation system ultimately leads to a misrepresentation of a commander's overall performance.* Therefore, to understand readiness, one must understand the individual training standards that govern it. These training standards reside as part of the training and readiness program.

1. The Training and Readiness Program

The Marine Corps T&R program has become the principal framework that standardizes how training is conducted and evaluated in the USMC (Marine Corps, 2011). The T&R program was pioneered by the aviation community in the mid-1970s to provide a standardized system for conducting training (Marine Corps, 2011). As training and maintenance costs for aviators increased, the community needed standards that qualified an aviator as being trained. Without such a system, individual commanders defined what it meant to be a combat capable aviator. This lack of standardization led to inefficient and sometimes dangerous training (Marine Corps, 2011). Once the aviation community established a standard skillset, formal evaluations were created to determine unit proficiency. This process of establishing a set of training standards, linking those standards to the concept of tactical proficiency, training them, and evaluating performance of those standards became the model for preparing all Marine Corps units for war.

As the 1970s ended, training standards were established for every occupational specialty to define the basic skills required for every Marine (Marine Corps, 2011). This process was optimized throughout the 1990s by adopting the *System's Approach to Training* (SAT) and *Unit Training Management Guide* (UTM). T&R standards are the backbone that provide commanders with standardized training outlines for all occupations (Marine Corps, 2011). Training standards provide commanders with a cohesive set of criteria to build combat capable units.

2. The Training and Readiness Task—Fire and Movement as an Example

The T&R task is made up of multiple parts which support organization, evaluation, training, policy, and support requirements. The task we are examining is "INF-MAN-3001: Conduct Fire and Movement." The task requires a unit of two Marines, an order to attack an enemy position, and the context of a larger unit to complete. The two Marines alternate shooting and moving to close with the target and neutralize the enemy. One Marine shoots to allow their buddy to move, and once that buddy finds cover, they provide fire to allow the first Marine to move. This process is outlined in detail in the performance steps, which must be executed to the standard. This thesis is focused on the standard portion of the T&R task (highlighted in red in Figure 1) which dictates the level of performance an individual should display in the execution of a task. A sample T&R task is shown in Figure 1 followed by a discussion of the standard portion. For an in-depth description of every section of the training standard, refer to Appendix A.

INF-MAN-3001: Conduct fire and movement

SUPPORTED MET(S):
MCT 1.14 MCT 1.6.1 MCT 1.6.4

EVALUATION-CODED: NO **SUSTAINMENT INTERVAL**: 6 months

CONDITION: Given an order from higher and an enemy.

STANDARD: To neutralize the enemy threat in order to accomplish the mission, meeting the commander's intent.

EVENT COMPONENTS:
1. Suppress the enemy (S).
2. Assess effects of fires (A).
3. Adjust fires as necessary.
4. Identify next covered position.
5. Move to next covered position under the cover of suppression(M).
6. Identify your target and continue suppression to allow buddy to move to next covered position.
7. Repeat steps 1-5 until the objective is reached.
8. Execute actions on the objective (K).
9. Consolidate.

Figure 1. INF-MAN-3001 highlights a lack of quantifiable standards. Source: Marine Corps (2016a).

The Standard: The standard is the focus of this thesis and "indicates the basis for judging the effectiveness of the performance. It consists of a carefully worded statement that identifies the proficiency level expected when the task is performed" (Marine Corps, 2011, p. 4–3). The standard provides the lowest level of performance to qualify as trained in the given task and can range from specific quantitative metrics for individual events to general statements for collective events. The example standard in Figure 1, "to accomplish the mission and meet commander's intent," provides no quantifiable metrics or measures of performance and calls into question the Marine Corps' statement that "the standard" is a "carefully worded statement that identifies the proficiency level expected" (Marine Corps, 2011, p. 4–3).

The meticulous organization of T&R tasks shown in Figure 1 is one example that represents the Marine Corps' analytical approach to conducting initial, sustainment and combat preparation training. Unfortunately, regardless of how structured the framework is, the standard allows for a wide degree of immeasurable and acceptable performance. This elasticity makes comparing performance or trying to understand performance extremely difficult.

C. CURRENT CAPABILITY GAPS IN MILITARY EVALUATIONS AND STANDARDS

The T&R task dictates how tasks should be trained and reports how prepared a unit is for combat. However, the current model for evaluations does not afford evaluators the ability to describe performance beyond the binary "Go/No Go." Standards are spelled out for each T&R task, but these standards rarely provide quantitative means for evaluating the task known as measures of performance (MOP). This lack of clarity forces evaluators to generalize performance using measures of effectiveness (MOE). Due to the vast breadth of performance steps the standards cover, it is often difficult to establish quantitative MOPs for the overall task. Judgement, built on experience, must be utilized in order to determine proficiency using MOEs. The Marine Corps needs to institute a method to evaluate task performance quantitively as well as qualitatively. Quantifiable MOPs for all tasks could provide the proof necessary to determine the effectiveness of training programs and systems.

The difficulty in quantifying task performance with MOPs is that not every task has clear and quantifiable standards. For example, the individual task "0300-RFL-1003: Zero the Weapon" has the standard, "Achieve 3 out of 5 shots within a 4 minute of angle group at a specific range" (Marine Corps, 2016a, p. 8–33). The number of rounds in a 4 minute of angle group serves as a quantifiable MOP for evaluating performance. However, the previously referenced collective task "INF-MAN-3001: Conduct Fire and Movement" has the standard, "Neutralize the enemy threat in order to accomplish the mission, meeting commander's intent" (Marine Corps, 2016a, p. 7–56). In contrast to the first task, this task does not have a quantifiable standard that can be used for evaluating performance. Instead, a buddy pair of Marines is considered trained in fire and movement if the evaluator decides that the pair accomplished the performance steps according to the MOE outlined in the standard.

This begs the question, is it possible for one buddy pair to complete a task more effectively than another buddy pair? If the answer is yes, then the MOE outlined in the standard serves as a minimum acceptable measure of success that pairs are capable of improving upon. As stated in Douglas Hubbard's book, *How to Measure Anything*, "if it matters it is observable, if it is observable, it can be detected in an amount, and if it can be detected in an amount it can be measured" (Hubbard, 2014, p. 39). The trouble is that there are thousands of tasks, and each task has a different measure of performance. We need a measurement system that is capable of plugging into the individual training standard framework and allows evaluators the opportunity to quantify performance.

D. CONCLUSION

Today, USMC evaluators are trained during a multi-phased training and education process to identify what optimal performance looks like in each task, but they are not always given accurate MOPs to evaluate performance quantitatively. Standardizing evaluations allows evaluators to consistently measure task performance. Today the after-action review provides exhaustive qualitative feedback to trainees and units, but that feedback is reduced to a binary score that loses the granularity required for comparing performance. As a result, units cannot compare their performance in tasks to their peers or

to the organizational average. A better approach for measuring performance would be to provide evaluators with the ability to quantify the performance of the task using a metric which fits within the established T&R program. This metric would leverage the power of the T&R task, while adding a quantifiable MOP. The next chapter consists of a literature review that guided the development of SPEMS to address the aforementioned issues and following research question.

E. RESEARCH QUESTION AND HYPOTHESES

The following research question was carefully composed to guide the research described in this thesis. The research question is accompanied with three evaluation criteria that detail how to determine if the research question was answered entirely or partially.

1. Research Question

Can we define, develop, and assess a SPEMS that quantifies individual and collective performance in a specific task that is more effective than the current binary measurement, and shows potential for working uniformly across all tasks and missions?

a. Evaluation Criteria 1

SPEMS definition: We defined SPEMS based on current best practices that are internally reviewed and accepted as capable of accurately measuring task performance. Behavioral anchors were developed and validated using card sorting techniques wherein focus groups sorted and grouped anchors according to performance level. Anchors were retained based on focus group agreement.

b. Evaluation Criteria 2

SPEMS development: We incorporated SPEMS into the training evaluation tool for an experiment. This step included the conduct of a pilot study and was accompanied with a survey of subject matter experts. Acceptable levels of inter-rater reliability were reached in terms of how different subject matter experts evaluated the same behavior using SPEMS. Additionally, the subject matter experts decided that the objective MOP for the buddy rush was the percent a buddy pair is exposed during the task.

(i) Cronbach's Alpha > 0.70

c. Evaluation Criteria 3

Assessment of SPEMS– We conducted live experimentation to compare current performance evaluation methods and SPEMS to determine which evaluation measure provides better information about the proficiency of individuals and groups of Marines.

2. Hypotheses

a. Hypothesis 1

- H_0: There is no relationship between SPEMS scores and objective measures of performance.
- H_A: There is a relationship between SPEMS scores and objective measures of performance.

b. Hypothesis 2

- H_0: There is no difference in the predictive strength between SPEMS scores and PECL on objective measures of performance.
- H_A: There is a difference in the predictive strength between SPEMS scores and PECL on objective measures of performance.

THIS PAGE INTENTIONALLY LEFT BLANK

III. BACKGROUND

A. INTRODUCTION

Measuring the benefits of training is necessary to determine the levels of effectiveness and efficiency of the training. The military has a very specific way of training; understanding the effectiveness of training is important for many reasons. For programs like sales training, identifying the benefits are easily done by measuring and comparing the upfront costs of the training to the measured revenue benefits (Hubbard, 2014). Often it is difficult to tease out benefits for more obscure training programs; however, the field of training effectiveness and performance measurement has developed methods like Kirkpatrick's four levels of learning evaluation (Kirkpatrick & Kirkpatrick, 2006). We need to leverage these proven methods to provide quantitative feedback to overcome the difficulties associated with measuring military task performance.

Outside of actual combat, every aspect of military day-to-day activities is comprised of either training or operations. Here we are focused on training. The Marine Corps offers countless examples of training programs that were developed using a system's approach to training (SAT), managed using the Unit Training Management Guide (UTM) and executed according to the publication *How We Train*. We examine these publications in the first section of this chapter. These publications provide an extremely strong framework for designing and executing training but, as noted in Chapter I, they lack a definitive method for evaluating performance.

Consequently, we examine a study that designed guiding principles for the development of successful, measurable military training. These principles illuminate problems with current USMC training methods and propose how training can be improved. Subject matter experts are often required to be a part of training evaluation due to the nature and complexity of the tasks being evaluated. One challenge with using subject matter experts is the effect of rater bias on the reliability of evaluations.

In this chapter, we also address a common pitfall of performance evaluation systems, reliability and bias. After exposing these biases as a potential problem, we review

how to guard against them to ensure reliable evaluations are being conducted. After addressing this problem, we discuss a variety of training evaluation tools. Behaviorally anchored rating scales (BARS) is one tool that has been used outside of the military to evaluate performance quantitatively.

Finally, with an understanding of how to create a performance evaluation measurement system, we examine one proof of concept study that demonstrates how a performance evaluation tool that uses BARS could be used to assess a training system's effectiveness. This chapter provides the reader with a roadmap that explains both our research methodology, and how the research fits within the context of the larger problem.

B. MILITARY TRAINING: A PRACTICE AND AN EVALUATION

Training is a tool that is used by the military in order to instill discipline, evaluate tactical preparedness for combat, and practice complex processes. Education is the process of receiving instruction and knowledge for the sake of expanding one's mind. Training contains education but adds a component of practice for the purpose of skill acquisition. Formal military training in the United States was first published in 1807 by General Friedrich Wilhelm Von Steuben, a Prussian and American Army officer during the revolutionary war (United States War Department, Fleury, & Steuben, 1807). Von Steuben's *Regulations for the Order and Discipline of the Troops of the United States* established training guidelines and the direct military tasks required by every troop in the Continental Army to prepare for war. Military training is the foundation that prepares individuals to risk their lives by conducting physically and mentally challenging missions (Fletcher & Chatelier, 2000b). The consequences of combat lead military and political leaders to invest a great deal of resources, into military training. To ensure forces are properly trained, the military must *evaluate* the individual and collective levels of training, as well as the training systems used to prepare them for combat.

An evaluation "determine [s] the importance, size, or value of" a process by an individual or organization ("Evaluation." n.d.). In educational terms an evaluation is a single process which passes judgement according to standards, goals, and criteria (e.g., a test) (Taras, 2005). The Marine Corps follows traditional educational models of evaluation

that utilize both formative and summative approaches to testing (Richardson, 2013). The summative evaluation sums the proof of, in this case, a student's importance, size, or value up to a given point (Taras, 2005). The formative evaluation is similar; however, for an evaluation to be formative it must also provide feedback (Taras, 2005). This feedback focuses on the difference between the work that was done, and the work that is required, and is typically captured in an after-action review.

The Marine Corps primarily utilizes a formative evaluation technique to both judge task completion and provide focused feedback. The evaluation process is outlined in the System Approach to Training (SAT) Manual which applies standard systems engineering to instructional programs in order to provide methodically derived instructional system designs (Marine Corps, 2004).

1. The Systems Approach to Training

The systems approach to training (SAT) grew out of the Department of Defense's (DoD) need to rapidly and effectively design training systems (Fletcher & Chatelier, 2000). To build this capability the Marine Corps sought to apply a systems approach to the process of building instruction. The model known as Instructional System Design (ISD) was instituted across the Marine Corps as the SAT. The SAT continues to serve as the primary method which governs the training development process (Marine Corps, 2004).

The SAT consists of five phases: analysis, design, development, implementation, and evaluation, (ADDIE, Figure 2), which serve as the guidelines for the development of training and instruction.

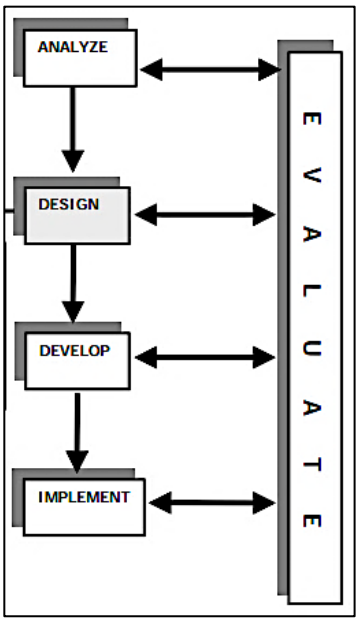


Figure 2. The ADDIE model. Source: Marine Corps (2004).

The SAT is a cyclical process where each phase builds upon prior phases by dynamically processing inputs and developing outputs (Marine Corps, 2004). Problem analysis consists of determining what skills or abilities are required to perform a specific job. A task analysis is used to convert these skills into ITSs which include six parts: the task, condition, standard, performance steps, administrative instructions, and references (Marine Corps, 2004). As stated in Chapter I, ITSs became the foundation of how every single Marine is trained through the development of the training and readiness (T&R) manual. The design phase converts tasks into learning objectives. The development phase converts learning objectives into periods of instruction. The implement phase is where the instruction takes place. Finally, the instruction is evaluated for its level of effectiveness during the evaluation phase (Marine Corps, 2004). This evaluation typically includes surveys, course critiques, and qualitative feedback, but does not measure task performance quantitatively (Fletcher & Chatelier, 2000). The SAT manual provides a detailed approach for developing formal training and periods of instruction, but the management and application of those training plans is not the focus of the publication. The UTM provides specific instructions on how military units should train and evaluate performance of collective tasks.

2. The Unit Training Management Guide

History, experience, and wisdom have proven that there is a direct relationship between training and victory in war (Marine Corps, 2016c). The unit training management guide (UTM) was written to assist Marine Corps units in the development of unit training programs. The UTM outlines a training philosophy that mandates the Marine Corps provide combat ready units to the nation (Marine Corps, 2011). To meet this mandate, the Corps outlined fundamental training principles in the UTM. These training principles are designed to guide the way units conduct and build training programs. We focus on two major principles of the UTM, "Use standards-based training," and "Use performance-oriented training" (Marine Corps, 2016c, p. 1–3).

The USMC currently uses a binary mechanism for determining if tasks were trained, or untrained, depending on whether or not the task was performed to a given standard (Marine Corps, 2016c). A task is defined as, "a unit of work usually performed over a short period of time. A task has a specific beginning and ending, can be measured, and is a logical and necessary unit of performance" (Marine Corps, 2016c, p. 4–3). A standard is defined as, "accuracy, time limits, sequencing, quality, product, process, restrictions, etc., that indicate how well a task should be performed. Simply stated, a measure of performance" (Marine Corps, 2016c, p. 4–3). For tasks that are not easily quantifiable, meeting the standard is typically measured by observing if the completion of a task involved the implementation of each performance step. If the performance steps were all accomplished in accordance with the standard, then a Marine is considered to be trained in that task. These ITSs become the foundation of not only what needs to be trained in each specific Marine occupational specialty (MOS), but also how each task is to be performed through the establishment of performance steps, conditions, and standards. As ITSs became T&R events, events were assigned to different training locations and times which naturally turned into standard training cycles.

Different T&R events are used to build different parts of the training cycle based on how they are coded. Formal schools are responsible for training the "F" coded individual events in their initial training setting. Operational units are responsible for maintaining the proficiency of events according to their sustainment interval, and training new events based

on how they support the commander's METL. One instance of formal schooling is the School of Infantry which develops all enlisted infantry coded (03xx) MOSs. This cycle of expanding proficiency in formal schools, developing experience and sustaining that proficiency in operational units, and evaluating combat effectiveness is how the Marine Corps trains to fight.

3. How the Marine Corps Trains to Fight

a. Education

Education is the foundation to any combat capable Marine that provides the background knowledge necessary for training. Education conducted in a classroom setting provides Marines with formal instruction designed to teach Marines how to perform specific tasks. Education can range from the professional military education (PME) program to periods of instruction designed to explain how a task is performed (Marine Corps, 2008).

The enlisted PME program focuses on developing military decision making and leadership that serves as the foundational philosophy from which all MOS specific skills are built upon (Marine Corps, 2008). This program begins with Marine Corps recruit training and continues on through the career of each Marine. PME is coupled with MOS specific education to build tactically proficient small unit leaders who are capable of making decisions. Once these skills have been introduced in the classroom environment, practical application of the skill provides a kinesthetic approach to learning that supports skill transfer (Marine Corps, 2004).

b. Formal Schools Training

"Core skills" are the MOS specific skills that are critical for a Marine to succeed in combat and required to become as a member of that MOS. Entry-level schools are mandated to train every MOS candidate in these skills to guarantee their proficiency prior to joining the operating forces. (Marine Corps, 2011). These skills are trained using the "crawl-walk-run" approach to training wherein Marines are educated on a given skill, given a chance to apply the skill in practice, and evaluated on the skill during evaluation events

(Trabun, 2007). Approaching training in this way is absolutely critical due to the "building block approach to training" employed by the T&R manual (Marine Corps, 2011).

The building block approach, also called serial part task training, is where Marines practice and are evaluated progressively more complex tasks in order to become ready for combat (Marine Corps, 2011). In theory, graduation from formal schools indicates that Marines have mastered their individual core skills and are prepared to support combat capable units immediately. This is often not the case in practice. Core skills are further developed and built upon during operational job training (OJT), where Marines are presented with additional skills and tasked with evolving their leadership ability. The training process is outlined more specifically in MCRP 3–0B "How to Conduct Training," which prescribes an educate, demonstrate, practice, and evaluate continuum for skill development (Marine Corps, 2015a).

c. Practical Application—Deliberate Practice

In the USMC, deliberate practice is called practical application. Practical application typically begins with a demonstration of the skill and finishes with students practicing the skill in a less dynamic environment. Practical application is a tool instructors use to coach and evaluate the progress of their students (Marine Corps, 2004). Practical application provides instructors with an informal evaluative means for developing a student's ability to complete a given ITS prior to live-fire qualification. This process is done to maintain safety, to ensure students are capable of conducting the task, and to optimize the use of training resources prior to moving to a dynamic and costly training environment (Marine Corps, 2015a).

d. Live Fire Qualification

Live fire qualification is conducting deliberate practice at the speed and under the conditions prescribed by the T&R task in order to achieve the standard specified (Marine Corps, 2015a). Drills are progressive in nature and are conducted with all equipment at an increased pace and under more realistic conditions. Performance is evaluated by the unit's leader or any commander in the chain of command of that unit. Deficiencies are identified and remediation practice is used to reconcile any gaps between Marine performance and

the T&R event's standard (Marine Corps, 2015a). If these T&R tasks are evaluation (E) coded then the unit's combat readiness percentage (CRP) is updated for the specific MET the task is associated with (Marine Corps, 2011).

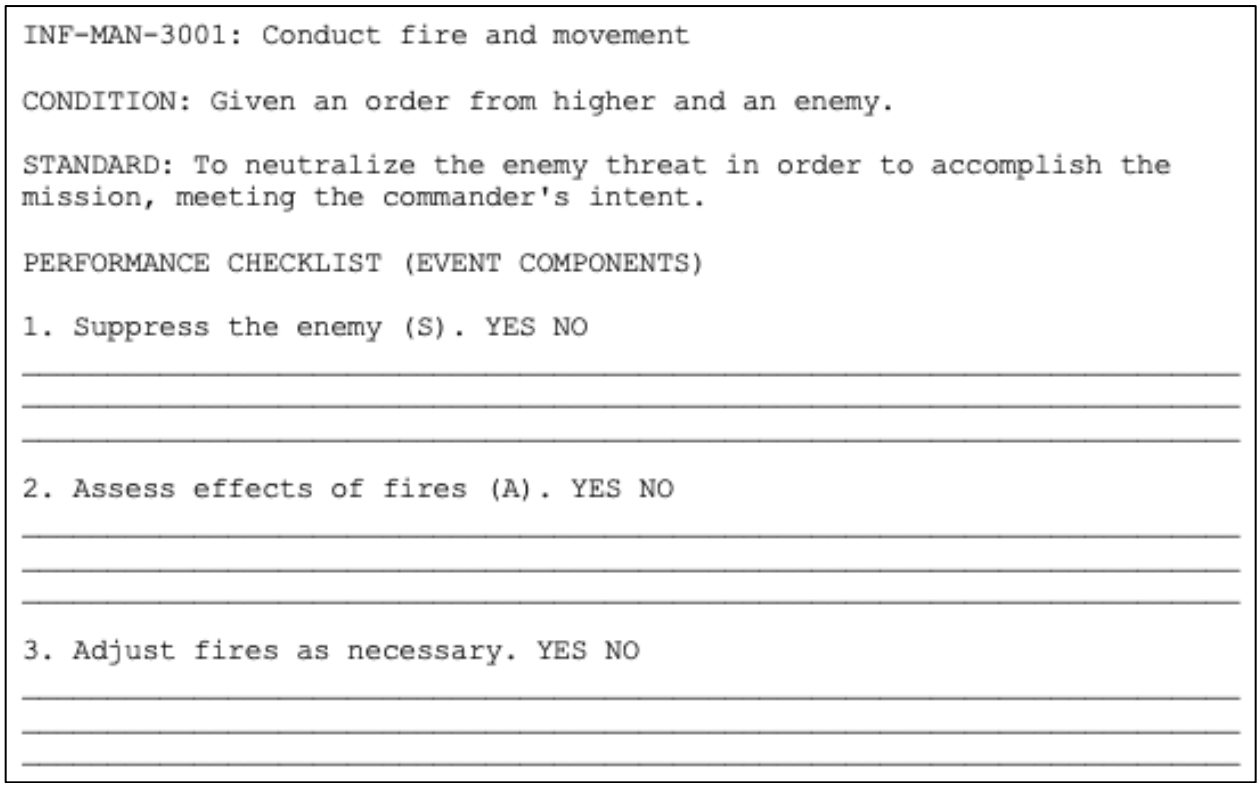
The unit's CRP increases as more E-coded events are determined to be "trained," which enables commanders to identify areas in which they are proficient and deficient (Marine Corps, 2011). As commanders approach deployment to operational theatres they seek to maximize their CRP in order to demonstrate their unit's level of readiness. The calculation of the CRP is based solely on the number of events evaluated as trained (Marine Corps, 2011). This is no guarantee as to the level of individual or unit proficiency in any of these events. The combination of the importance of the CRP and the lack of accurate quantitative performance assessment leads to a gross misunderstanding of readiness and task proficiency. Fortunately, the live fire qualification, evaluation, and remediation model does provide qualitative feedback in the form of an after-action review.

e. The After-Action Review

The after-action review (AAR) has been used for over 45 years as a formative evaluation that affords individuals and groups a means to correct deficiencies in training (Richardson, 2013). It is ostensibly a qualitative evaluation that serves as the primary feedback mechanism of the military to correct behavioral deficiencies in training (Morrison & Meliza, 1999). This subjective measurement of performance is primarily a process evaluation where experienced observers analyze the techniques used to achieve the T&R task's desired outcome.

The AAR is typically integrated with the aforementioned performance steps in a document known as a performance evaluation checklist (PECL) to determine whether or not the training audience is proficient in a given task (Marine Corps, 2015a). The PECL, previewed in Figure 3, highlights the specific performance steps that are required which informs the qualitative feedback necessary to get the training audience to address the gap. *Unfortunately, this process does not provide a quantitative evaluation of how well these performance steps were accomplished nor does it stratify performance beyond the binary*

measurement of being "Go/No Go"; herein lies the problem with how evaluation is currently conducted.



INF-MAN-3001: Conduct fire and movement

CONDITION: Given an order from higher and an enemy.

STANDARD: To neutralize the enemy threat in order to accomplish the mission, meeting the commander's intent.

PERFORMANCE CHECKLIST (EVENT COMPONENTS)

1. Suppress the enemy (S). YES NO

2. Assess effects of fires (A). YES NO

3. Adjust fires as necessary. YES NO

Note: Performance steps 4–9 omitted due to redundancy.

Figure 3. INF-MAN-3001—Performance evaluation checklist. Source: Marine Corps (2013).

The PECL needs improvement to provide Marines with proficiency measurements. Understanding military training has been the focus of a number of studies that were designed to help improve procedures.

C. A REVIEW OF BOLDOVICI'S MILITARY TRAINING ASSESSMENT AND IMPROVEMENTS

1. Overview

Current formative training evaluation requires the use of measures of performance (MOP) and measures of effectiveness (MOE) that provide reliable and valid scores to

capture task performance quantitatively. All U.S. military services utilize MOPs and MOEs for evaluating tactics, maneuver, combat attrition, and other modeling. These ratings are used for both training effectiveness, and system evaluation (Boldovici, Bessemer, & Bolton, 2002). Boldovici et al. address two kinds of ratings, analytic and performance-evaluation (Boldovici et al., 2002). We have discussed the problems associated with current formative evaluations in the USMC. In subsequent sections we demonstrate how new performance measures can be introduced to strengthen training evaluations. These new methods are substantiated by Boldovici et al.'s (2002) guiding principles for developing military training and evaluation.

2. The Guiding Principles for Developing Military Training and Evaluation

Boldovici et al. outlined several principles to be used as a way to properly evaluate training and training systems. The development of SPEMS was guided by these properties. The three essential properties of rating scales are: reliability, validity, and generality. The reliability of scores is important for ensuring ratings from a variety of raters is externally consistent (Gliner, Morgan, & Leech, 2011). Reliability is computed by estimating the consensus among raters regarding performance on a specific T&R task (Boldovici et al., 2002). The validity of scores speaks to how accurately the rating system captures performance of the task. Generality is achieved through an institutional commitment to score and collect scoring data over time to improve the generality of scores from single task estimates across the force. Applying a developed rating system institutionally should quickly demonstrate evidence for or against the system's ability to work for a variety of tasks. This thesis focuses on the properties of reliability and validity over the following paragraphs.

Reliable performance evaluation techniques should be added to current field-testing in order to provide reliable and valid performance data that adheres to Boldovici et al.'s (2002) essential properties of ratings. Current field trial scores are based on the ability of observer-controller's to boil down performance on tactical tasks to a binary measurement (Boldovici et al., 2002). This binary measurement reduces the reliability and maximum possible validity of field trial results; however, if proper inter-rater reliability, rater

preparation, behavioral anchoring, and recording techniques were applied, the Marine Corps could leverage existing field training performance evaluation techniques to reliably validate training systems and procedures.

With these properties in mind, there are three phases to designing a rating system for reliability and validity: rater preparation, observation, and recording (Boldovici et al., 2002). Raters should be standardized utilizing current train the trainer practices throughout the Marine Corps (Marine Corps, 2004). These training procedures should be strengthened through specific instructions, practice, feedback, and tests that show institutional evaluators how to utilize performance rating scales to evaluate task performance (Boldovici et al., 2002). Raters should then be allowed to consistently and continually rate similar events in order to broaden their understanding and calibrate their ability to evaluate performance (Hubbard, 2014). The combination of these techniques with scoring aids and templates has the potential to produce valid and reliable task scores. Scoring aids that define the task, the various levels of performance in the task, and how to properly score the task decrease the measurement error of the scorer (Boldovici et al., 2002).

The validity of rater evaluations can be improved by controlling observational variables and decreasing the cognitive load of the individual raters. Multi-dimensional tasks should be deconstructed into performance steps in order to yield more valid ratings for all tasks (Boldovici et al., 2002). The Marine Corps currently accomplishes this tenet by decomposing training and readiness tasks hierarchically into sub-tasks that are chained to larger tasks (Marine Corps, 2011). The lower level tasks are decomposed into performance steps that capture performance of a task according to its basic behaviors (Marine Corps, 2016a). Decomposing tasks into observable behaviors has the potential to increase validity and reliability when paired with anchored performance evaluation systems.

Boldovici et al. (2002) recommends stabilizing the observability of practice to limit the effect of visual noise and allow for more detailed analysis of performance. Current video play back technologies and performance evaluation tools provide technological solutions that could greatly enhance the rater's ability to avoid the effects of this variability

but are outside the scope of this work. Video technologies should be paired with new quantitative evaluation measures to improve the external consistency of evaluations.

Following the properties described by Boldovici (2002) for developing and implementing rating procedures has the potential to produce valid, reliable, and generalizable rating systems for measuring task performance, and ultimately training effectiveness.

D. THE RELIABILITY OF MILITARY RATERS

1. A Review of Inter-rater Reliability Pitfalls and Military Solutions

Evaluating complex and ambiguous situations using highly calibrated measurement instruments is often next to impossible (Hubbard, 2014). Instruments like thermometers can be used to measure objective qualities like temperature, but complex behavioral situations require the use of human beings. Unfortunately, the human mind is a complex system involving a number of biases that influence the human's ability to reliably judge behavior. Two such biases are anchoring (how being given a starting point affects people's estimates) and the halo effect (a rater scoring an attractive person's performance more positively than that of an unattractive person) (Brewer & Chapman, 2002; Tversky & Kahneman, 1974).

In most estimates of anything, people make comparisons between their starting point, or initial value, and what they are observing (Brewer & Chapman, 2002; Tversky & Kahneman, 1974). An estimate is made up of a starting point plus how what the person is observing differs from that starting point. The starting point plays a large role on the estimate and varies from person to person. Different starting points yield different estimates leading to what is known as anchoring bias (Hubbard, 2014).

One example of this phenomenon was demonstrated by Tversky and Kahneman when one group of subjects was asked if the African percentage of the United Nations was greater than 10% and another group was asked if it was less than 65%. Both groups were then asked to estimate the percentage and the critical finding was that the second group whose estimate was anchored at 65% gave an answer 20% higher than the other group (Hubbard, 2014). This study shows the effect of anchoring bias on a person's ability to

estimate. Tversky and Kahneman goes on to explain that a judge must be properly calibrated in order to provide accurate and reliable measurements (Brewer & Chapman, 2002; Tversky & Kahneman, 1974), and this sentiment is further supported by Hubbard's emphasis on rater calibration (Hubbard, 2014).

The military spends significant resources, on training, educating, demonstrating, imitating, and practicing correct behaviors in the performance of tasks to both build task proficiency and rater reliability. This training can be leaned on to build a common baseline that reduces the effects of anchoring bias, but there are other biases and heuristics that need to be guarded against.

The halo effect is the idea that if a rater favors or disfavors certain attributes of a ratee than the rater can be predisposed to interpret the ratee's performance in accordance with their attribute conclusion (Hubbard, 2014). For example, if a Marine is well received by the unit, then, it is more difficult to rate that Marine poorly in the execution their duties, regardless of the actual performance of the duties.

This bias was studied by Robert Kaplan of San Diego State University who demonstrated the positive correlation between the attractiveness of a person and their grade on an essay. The exact same essay was assigned to a number of graders, and the picture attached to it was randomly assigned. Higher grades were strongly positively correlated with subject attractiveness which demonstrated evidence of the halo effect (Kaplan, 1978).

This type of bias is typically guarded against through the use of calibration (Hubbard, 2014). By anchoring the rater's evaluations in behaviors, the organization relieves the rater of the responsibility of being the judge. Raters are instead able to act as observers who match behaviors with scores, thereby limiting the influence of external biases (Schwab, Heneman Herbert G., & DeCotiis, 1975). These biases are difficult to guard against due to their presence inside human decision-making and judgement; however, Wigdor and Green's *Performance Assessment for the Workplace: Volume 1* provides evidence that military evaluators are properly calibrated to guard against these issues (Wigdor & Green, 1991).

2. A Review of Performance Assessment in the Workplace: Volume 1

The Joint-Service Job Performance Measurement Standards (JPM) project was an effort to develop measures of performance in entry-level military jobs in order to link on-the-job performance with recruitment standards (Wigdor & Green, 1991). The project was initiated by Congress in the mid-1970s when the abolishment of military conscription gave way to the prospect of maintaining an able-bodied all-volunteer force. The project first examined psychological and intellectual tests finding their irrelevance compared to measuring individual service member proficiency.

After determining intellectual tests provided minimal predicative ability, Wigdor and Green (1991) performed a detailed literary examination of job performance criteria such as absenteeism or accident rates to highlight how these types of metrics led to the criterion problem. The criterion problem describes the relative inability of a wide spectrum of criterion to accurately measure work performance. This phenomenon was summarized by Landy and Farr who concluded that accurate measurement of performance lies beyond an individual criteria's ability to define a person's proficiency (Landy & Farr, 1983). Instead, performance evaluation techniques are suggested as possible ways to understand and evaluate task performance. As long as rating reliability and validity are both measured and controlled, human raters are a viable option for measuring task proficiency.

Proficiency, defined as a person's advancement of skill to the state of being proficient ("Proficiency," n.d.), was observed through the selection of hands-on performance tests or work samples in order to faithfully benchmark job performance in daily tasks (Wigdor & Green, 1991). The JPM project performed a comprehensive study that combined the results of multiple experiments in an effort to capture military on the job performance. The project was able to lean on the aforementioned task analysis done by the Marine Corps which created the individual training standard (ITS) as part of the development of instruction system design (ISD) in the 1970s (Fishburne, Murray, & Blair, 1979). These ITSs became the T&R tasks that defined the facets of performance measured by Wigdor and Green. Wigdor and Green concluded that military jobs can be accurately modeled as a collection of tasks, which can therefore be measured to determine job performance (Wigdor & Green, 1991).

Thirty jobs with 15 tasks per job were analyzed using work samples, interview, simulations, multiple choice knowledge tests, and a variety of performance evaluation ratings (Wigdor & Green, 1991). Work samples were evaluated by having a Marine conduct an instance of a real task and scoring each behavioral step as either a *"Go/No Go"* by a trained observer. This evaluation is consistent with the Marine Corps' current policy on task evaluation (Marine Corps, 2015a). Utilizing observers introduced error and this error had to be assessed using a standard practice known as generalizability theory (Rubin, Cronbach, Gleser, Nanda, & Rajaratnam, 1974). When this error was analyzed, researchers saw that there was virtually no effect of the observer on the performance appraisal's stability (Wigdor & Green, 1991).

The reliability of any rating instrument can be observed through a variety of statistical tools and methods; Wigdor and Green utilized stability tests known as generalizability theory (G-Theory) and pairwise correlation to demonstrate a strong agreement between military raters (Wigdor & Green, 1991). The term generalizability refers to the environment in which the task was conducted. The observed score is decomposed into a universal score, or true score, and error components that are associated with each element of variation (Rubin et al., 1974). Each of the errors and interaction effects have underlying variances which can be analyzed using standard analysis of variance (ANOVA) techniques to determine the amount of systematic error associated with each effect. By demonstrating a minimal amount of systematic error associated with rater variation, Wigdor and Green (1991) were able to show high degrees of inter-rater reliability in the Marine Corps.

In the Marine Corps study, 150 infantrymen were asked to conduct multiple tasks with a total of 35 scorable units across two sites. Marines were evaluated by two examiners and their scores were tested for stability and correlation by using a G-theory analysis to determine the reliability of the scoring system. The chart in Figure 4 summarizes the results of the experiment.

Generalizability is the ratio of true to true plus error variance.
The true variance component is M in all cases.
The error variance components are:
Relative, 1*E*, 1*T*: $M \times E + M \times T + M \times E \times T$
Absolute, 1*E*, 1*T*: $M \times E + M \times T + M \times E \times T + E + T + E \times T$
Relative, 2*E*, 11*T*: $(M \times E)/2 + (M \times T)/11 + (M \times T \times E)/22$
Absolute, 2*E*, 11*T*: $(M \times E + E)/2 + (M \times T + T)/11 + (M \times E \times T + E \times T)/22$
SOURCE: Webb et al. (1989).

TABLE 6-4 Estimated Variance Components in Generalizability Study of Performance Test Scores for Marine Infantrymen, Replicated at Camps Lejeune and Pendleton

Source of Variation	Variance Components	
Lejeune		Pendleton
Marines (*M*)	11.69	9.13
Examiners (*E*)	0.00	0.00
Tasks (*T*)	33.05	35.56
$M \times E$	0.35	0.28
$M \times T$	72.91	67.38
$E \times T$	0.07	0.02
$M \times E \times T$	11.69	12.70
Generalizability Coefficients for 35 Tasks		
Relative	.81	.78
Absolute	.76	.72

Figure 4. Results and explanation of G-Theory analysis of 15 Marine infantryman. Source: Wigdor and Green (1991, p. 126).

Reliability for the 35-item test, for relative scores, was 0.83 for Camp Pendleton, and 0.80 for Camp Lejeune, demonstrating a startlingly high degree of agreement between raters (Wigdor & Green, 1991). All tasks were comprised of multiple steps in the standard "Go/No Go" format, and steps occurred in a predefined order according to the T&R task (Marine Corps, 2016a). Throughout multiple complex tests, the same findings emerged: raters did not appear to introduce measurement error due to the strategic development and selection of calibrated raters (Wigdor & Green, 1991). This study provides evidence that Marine evaluators are properly calibrated, anchored, and can be trusted to provide reliable observations of task performance.

This study serves as a foundational example that demonstrates that the current binary performance measurement system is a reliable measure of performance but lacks the granularity to evaluate performance across the spectrum of competencies. This study also provides a performance evaluation blueprint that was followed during our evaluation of SPEMS's. Wigdor's study provides significant evidence that the military contains the proper systems for calibrating and effectively employing human raters in order to evaluate

task performance. This assumption allowed researchers to leverage the proven rater-based performance evaluation methods that are examined in the following section to properly evaluate performance.

E. PROPOSED PROVEN TRAINING EVALUATION METHODS

1. Kirkpatrick's Four Levels of Training Effectiveness

One example of a tested user and rater-based approach to evaluating training is Kirkpatrick's Theory on the four levels of training effectiveness. Kirkpatrick developed his four levels of evaluation in 1959 when he published a series of articles titled, "Techniques for Evaluating Training Programs" (Kirkpatrick & Kirkpatrick, 2006). These articles outlined four iterative levels of evaluating the effectiveness of training programs. The four levels of training evaluation are: Level 1—Reaction, Level 2—Learning, Level 3—Behavior, and Level 4—Results (Kirkpatrick & Kirkpatrick, 2006). These four levels are as applicable to the military domain as they are in industry, and are applied with comparable frequency (Fletcher & Chatelier, 2000). These levels are often used to understand how to improve future training programs, and to justify the existence of current training practices (Kirkpatrick & Kirkpatrick, 2006). The following is a brief explanation of the Kirkpatrick levels of evaluation.

a. Level 1: Reaction

Evaluation in this level measures how training participants react to the training program (Kirkpatrick & Kirkpatrick, 2006). Evaluating this level is done by eliciting satisfaction that informs future training program improvements, provides quantitative baselines, and builds trust between trainees and instructors (Kirkpatrick & Kirkpatrick, 2006). This type of evaluation is normally conducted by having users' complete surveys in order to indicate how content trainees are with the training.

b. Level 2: Learning

Evaluation of learning is the determination of what knowledge was learned, what skills were developed, and what attitudes were improved (Kirkpatrick & Kirkpatrick, 2006). In the case of knowledge and attitude, evaluation can be done in the form of a written

examination, but skills must be evaluated using a performance test (Kirkpatrick & Kirkpatrick, 2006). It is important to point out that by measuring skill and knowledge acquisition, the Marine Corps is actually evaluating its own effectiveness to instruct core competencies (Kirkpatrick & Kirkpatrick, 2006). These types of evaluations are occasionally implemented in military training as end-of-course measurements of skills; however, these evaluations are rarely quantified beyond a pass or fail determination (Fletcher & Chatelier, 2000). This makes it more difficult for the military to move from level two to level three because no change in behavior can be expected unless there was a measurable change in learning (Kirkpatrick & Kirkpatrick, 2006).

c. Level 3: Behavior

Evaluating behavior is a diligent accountability process that attempts to determine how much classroom learning actually transfers to the operational environment (Kirkpatrick & Kirkpatrick, 2006). There are few if any evaluations that measure on the job performance improvements quantitatively (Fletcher & Chatelier, 2000). The fitness report is an anchored, qualitative evaluation that attempts to measure a Marine comprehensively; however, this report is not tied directly to the outcomes of training. The report is primarily designed to support selecting, promoting, and retaining the most qualified Marines (Marine Corps, 2015). Therefore, the report builds a qualitative description of a Marine's value but does not evaluate behavioral shifts or performance in specific tasks as a result of that Marine's training. This is because the Marine Corps lacks quantifiable measures of performance that can be measured as a result of potential behavioral changes.

d. Level 4: Results

Evaluating results is defined as tabulating the benefits that occurred because participants undertook the training program (Kirkpatrick & Kirkpatrick, 2006). This step may be the most difficult in the context of this thesis due to the dynamic and multifaceted nature of military success. Industry typically examines issues of productivity and profitability whereas the military attempts to understand the relative combat effectiveness improvements as a result of training (Fletcher & Chatelier, 2000). There is both a scarce

and inconsistent set of historical examples of level four evaluations in the military. These evaluations are often domain specific and serve to justify a programs existence. This dearth of level four evaluations provides more evidence that the USMC needs to develop generalizable methods for quantifying combat effectiveness.

e. Conclusion

Most military evaluations do not proceed past Level 1 (Fletcher & Chatelier, 2000). Surveys, end of course critiques, and instructor evaluations serve as the primary evaluation tool for understanding the effectiveness of various formal military training programs; however, some programs do advance to level 2. The reason that so few evaluations are able to move along this continuum is a lack of quantifiable results. Behaviorally anchored rating scales (BARS) are one method that has the potential to leverage the reliability of military raters to score tasks and thereby determine the effectiveness of a training system on improving task performance.

2. An Introduction and Review of Behaviorally Anchored Rating Scales

Behavioral measures are defined as, "performance dimensions and scale values in behavioral terms" (Schwab, Heneman Herbert G., & DeCotiis, 1975, p. 550). Behaviorally anchored rating scales (BARS) are one example of behavior measures which have grown in popularity since their development in 1963. BARS provide an interesting alternative to traditional graphic rating scales (e.g., below average, average, above average) because they theoretically reduce the number of judgements the rater needs to make about the ratee (Schwab et al., 1975). Raters act more in the role of observers, and the inferential requirements to judge task performance are left to those who develop the BARS.

BARS are developed using an iterative process where subject matter experts provide the critical behaviors associated with the task, group these behaviors by expertise level, and rate the behaviors associated with each expertise level on their ability to represent the level of performance (Schwab et al., 1975). The first step, Critical behavior elicitation, is the process of developing the range of behaviors that could happen in a given task. These behaviors are then organized under a performance level (5, 7, or 9-point scales are common) so that the behaviors under each level match the score. Next, a second group of

independent raters are asked to bunch the behaviors by expertise level to determine agreement. Typically, a behavior is retained if fifty to eighty percent of the second group assigns it to the same expertise level as the first group (Schwab et al., 1975). Finally, each performance level is assigned a score based on the degree to which the behaviors capture effective performance on that level. A standard deviation maximum is established, and only behaviors that are rated with smaller standard deviation than the maximum are retained (Schwab et al., 1975). This process yields the overall BARS for the specific task and should be completed for each task in the training and readiness program to achieve the most reliable set of BARS. Figure 5 illustrates a final version of a standard BARS.

While the standard BARS in Figure 5 provides more detailed behaviors, the task decomposition done to create the T&R task could be levied to remove the need to develop specific BARS for each task.

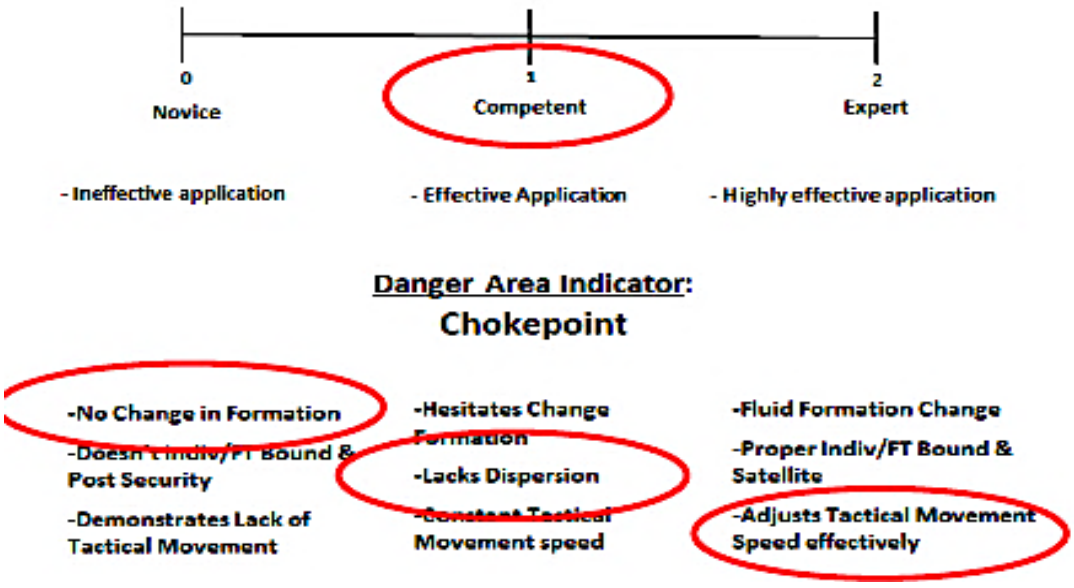


Figure 5. Standard BARS with selected behaviors.
Source: Richardson (2013. p. 77).

The T&R task establishes a list of performance steps which serve as the critical incidents necessary for completing the overall task (Marine Corps, 2016a). In order to develop BARS associated with a training task, numerical rating scales could be applied to

these validated performance steps to determine the level of expertise required in the performance step. Applying numerical rating scales would maximize existing work conducted by the Marine Corps to define, develop, and validate the necessary behaviors to achieve success in a given task. However, simply asking raters to determine how well a ratee conducts a performance step on a numerical scale does not provide enough reliability (Kingstrom & Bass, 1981).

General behavioral characteristics need to be attached to the rating of each performance step in order to anchor raters' scores effectively. The challenge is to select verbiage that leverages the reliability of BARS instead of the leniency of graphic ratings (Schwab et al., 1975). Verbiage of the anchors must be interoperable with a wide array of established T&R performance steps, while providing specific enough cues to inform raters as observers rather than as judges of performance. A vast array of studies demonstrate the power of BARS to yield very high reliabilities amongst raters (Schwab et al., 1975), but it is critical that performance step incidents are properly and generally described at each level in order to take advantage of this property. Leveraging the power of BARS to quantify task performance provides the potential for defining the value of a training system by its ability to improve a Marine's proficiency in a task quantitatively. The quantitative benefits of training using specific systems could be compared to determine the most efficient way to train tasks. This relationship could ultimately come to define the overall training value of an alternative training system, like the one described in the following proof of concept studies.

F. MOVING FROM FEATURES TO PROFICIENCY—DETERMINING A TRAINING SYSTEM'S VALUE AS A RESULT OF PERFORMANCE DATA: A PROOF OF CONCEPT

Performance evaluation and analytic ratings exist to provide different viewpoints on the training capabilities provided by a training system. Performance evaluation ratings are applied to Marines' collective and individual performance on tasks during field trials rather than to the devices themselves (Boldovici et al., 2002). Performance evaluations focus on how well the individual performs the task and are the focus of this study. Unfortunately, The Marine Corps currently uses training effectiveness evaluations to

conduct analytic ratings that attempt to leverage subject matter experts (SME) to conduct a total system evaluation.

Analytic ratings take the form of total system evaluation that involves eliciting SME judgement regarding the system's ability to train tasks. The Marine Corps uses a Systematic Team Assessment of Readiness Training (START) process to confirm the capability of systems to provide value added training in a distributed virtual environment (Dunne et al., 2017). The problem with current analytic rating systems is that they measure a system's capability to train specific tasks, but do not verify the system's ability to increase performance in those tasks. "The results of analytic evaluations applied to date have been unsuccessful in estimating training transfer" (Boldovici et al., 2002, p. III-5). Dunne et al. (2014) provide an example where performance evaluation ratings were successfully used to demonstrate the capability of systems.

Dunne et al. (2014) evaluated a simulation-based training system by examining two factors: the system's contributions to trainee performance and the costs avoided by using the simulated training system. They conducted this study in response to the Government Accountability Office Report 13–698 titled *Better Performance and Cost Data Needed to More Fully Assess Simulation-Based Efforts*, which states that the Services, "lack key performance and cost information that would enhance their ability to determine the optimal mix of training and prioritize related investments" (Government Accountability Office, 2013, p. 2).

1. A Performance Evaluation Rating System Proof of Concept Study

Dunne et al. (2014) examined a group of representative tank crews utilizing the M1A1 Advanced Gunnery Training System (AGTS) simulator by monitoring a practice sequence of 10 gunnery table tasks, with over 500 task instances, which culminated in a live-fire evaluation. The critical finding in Dunne et al. (2014) was that with, "*performance-oriented metrics and measures*, tied to doctrine and captured automatically, it is possible to determine both proficiency and cost avoidance" (Dunne, Cooley, & Gordon, 2014, p. 11). This research illustrates that quantitative performance metrics and measures are the critical element missing from current USMC training practices.

Dunne et al.'s (2014) study focused on the AGTS simulator. Each M1A1 AGTS GT VI exercise is composed of 10 collective tasks scored within a range of 0–100 with the passing standard being an average of 70 out of 100 across the 10 tasks. The tank community uses this scoring method as both a widely accepted and valid procedure for measuring a tank crew's ability to conduct gunnery and crew related tasks. Scores are gathered by the Heavy Brigade Combat Team (HBCT) using matrices that capture the target type, the posture of the tank, the range to the target, and the kill time (Dunne et al., 2014, p. 2). Typically, "this data is collected manually by the Instructor Operator (IO) who prints out the Performance Analysis, Situation Monitor, and Qualification Performance Analysis reports used for After Action Review (AAR)" (Dunne et al., 2014, p. 3). Table 1 depicts the average simulator and live-fire qualification scores for the three crews from Dunne et al. (2014), from the beginning of the training event to the end. These scores quantitatively depict the benefits of simulation-based training.

Table 1. Average simulator and live-fire scores for tasks by crew.
Source: Dunne et al. (2014, p. 7).

Crew	Average Beginning AGTS Scores*	Average Ending AGTS Scores*	Δ = Ending – Beginning Scores	Average Live-Fire Qualification Score*
1	63.6	93.0	29.4 (46% Increase)	90.7 (Result: Qualified)
2	55.2	81.9	26.7 (48% Increase)	85.0 (Result: Qualified)
3	53.6	87.0	33.4 (62% Increase)	78.0 (Result: Qualified)

* In conjunction with other evaluation requirements, a passing score is 70.

Each crew increased their average AGTS score by a minimum of 46%. This improvement could be attributed to more experience in the simulated training environment; however, examining the live fire qualification results in Table 1 shows that all crews passed their live-fire qualification with a close to equivalent score to their final AGTS score.

This result provides evidence that task proficiency achieved in the AGTS can transfer to the live-fire qualification. Furthermore, "trends in the proof of concept study indicate that task scores provided by the AGTS and from live-fire qualification, are appropriate reflections of performance for use in conducting proficiency studies…This

methodology could also be applied to other USMC training systems that *have similar task scoring systems*" (Dunne et al., 2014, p. 8).

Unfortunately, outside of the tank community *no such task scoring system exists*. However, by developing one we could demonstrate analogous cost-avoidance results like those depicted in Table 2.

Table 2. Number of simulated rounds fired by crew.
Source: Dunne et al. (2014, p. 7).

Crew	Munition 1	Munition 2	Munition 3	Munition 4	Munition 5	Munition 6	Totals
1	729	748	7	17	21	110	1632
2	460	1445	4	8	34	75	2026
3	999	877	29	36	48	176	2165
Subtotals	2188	3070	40	61	103	361	5823
Cost	$2,079	$10,684	$107,037	$163,232	$275,621	$966,011	$1,524,663

Table 2 shows the number of simulated rounds that were fired in the AGTS by crew in order to achieve the previously stated proficiency gains. This analysis does not consider tank operation costs, but only factors in the number of rounds fired in simulation. The cost of simulator training for these three crews is $7,208, and the costs avoided by conducting simulation based training totaled $1,524,663, illustrating a net cost avoidance of $1,517,455 (Dunne et al., 2014, p. 8). Performance data is the critical information necessary to answer the questions of whether or not to invest in different integrated LVC training environments. Performance data provides a quantitative method for comparing training outcomes. The performance data can be determined solely through the use of a quantitative task scoring system that allows comparison of performance when using different techniques and technologies. SPEMS has the potential to be one such system. Without a system like the SPEMS, the community can only speak about the number of rounds fired in simulation. If number of rounds fired in simulation is not positively correlated to an increase in proficiency, then the number of rounds 'saved' in simulation is irrelevant.

The tank community serves as a pioneer in metricizing task performance and should be used as an example of how the Marine Corps can measure task performance writ large. Dunne's proof of concept study provides an example of how SPEMS could be used to capture task proficiency, but training in the AGTS would have to be empirically compared to current training methods to determine the *training value* of training in the simulator.

2. The Importance of Performance Rating Evaluation on Determining Training Value

Jones et al. (2015) describe training value as the combination of a number of training related measures which include: "training task and performance capability, training realism capability, affective reaction level, and training efficiencies" (Jones, Seavers, Capriglione, & Jones, 2015, p. 3). Training efficiencies are the net costs of a system or the sum of costs and avoided costs. Training tasks and performance capabilities speak to the tasks that are able to be trained in simulation, and how well the tasks are performed in simulation. Current training effectiveness evaluations use analytic SME driven processes such as the START process (Dunne et al., 2017). These processes are used in order to determine what T&R tasks the system is capable of supporting. Currently, methodologies exist that allow us to gain insight into how well a training system can support the Marine's deliberate practice of tasks (training). One such example is the Integrated Training Environment Assessment Methodology (ITEAM) (Hodges, Darken, & McCauley, 2014). Missing from these methodologies are quantitative ratings of individual and team performance that allow for performance gains to be identified.

Performance data is the missing element necessary for determining the training value of the system. Performance data would allow for empirical side-by-side comparisons of existing and proposed (simulation) training solutions to determine the relative advantage of adopting a new system. SPEMS can provide the standardized performance measurement system for determining a Marine's task proficiency that results from training in each respective environment (Jones et al., 2015). By combining analytic training effectiveness evaluations (TEE) and cost avoidance data, with the comparative analysis, training capability developers could discount cost avoidance calculations to account for differences in the level of proficiency afforded by proposed training solutions. Discounting these costs

based on performance data would ensure proposed solutions return on investment was based on the training value they provided rather than merely ensuring avoided costs outweighed life-cycle costs.

The establishment of this training system evaluation plan addresses Jones et al.'s request for further research to, "establish standardized *training value* definitions and methods of analyzing factors to include cost, training effectiveness, and **efficacy** … TEEs and cost ROI analyses do not adequately address the cumulative value of training solutions" (Jones et al., 2015, p. 11). We need a performance measurement system that is capable of correlating existing analyses with performance improvements to ensure systems are acquired on the basis of training value.

G. SUMMARY

Training evaluation methods all typically involve some level of subjectivity; however, there exist methods for evaluating training that reduce rater bias, and quantitatively evaluate performance. Evaluating performance begins by understanding the tasks that are being performed, the training programs being used, and the implementation of those programs to enhance the effectiveness of Marines.

Starting from the base unit, the task, the Marine Corps developed the SAT as a way to standardize military training and allow units to develop and manage training reliably. The SAT is primarily a way to build training; however, as a part of the effort the Marine Corps conducted a thorough task decomposition of every major task. These tasks became the backbone to the training and readiness manual which is used to develop all training programs and schedules described in the Unit Training Management Guide, and How to Conduct Training publications. Training is built to satisfy readiness and currency requirements for every single task that is described in the task description itself. Unfortunately, the current binary methods for evaluating training do not satisfy Boldovici's (2002) guiding principles for developing and evaluating military training. The first step in meeting these principles is to determine that the evaluators that are necessary for evaluating tasks are properly baselined and calibrated to reduce bias.

We outlined common biases and pitfalls of using raters, as well as demonstrated how the military is properly set-up to use expert raters to evaluate performance. Anchoring bias, and the halo effect were offered as common examples of biases that are typically guarded against through proper rater calibration. By baselining and training to a standard set of tasks, Wigdor and Green (1991) demonstrated that military evaluators possess a high degree of inter-rater reliability which allows the military to use subject matter expert raters to evaluate training. Unfortunately, no method for quantitatively evaluating training performance currently exists. In order to solve this problem, we examined some methods for evaluating training.

We examined Kirkpatrick's four levels of training effectiveness and behaviorally anchored rating scales (BARS) as a possible way to evaluate Marine Corps task performance. Kirkpatrick's method proved to be unreliable and often incapable of providing anything more than qualitative feedback. The reason is a lack of quantifiable results in performance makes it difficult for training evaluations to proceed past level 1. In order to solve this problem, we examined BARS as a potential solution for providing quantifiable performance data resulting from training programs. Generalized BARS could be layered on the performance steps of T&R tasks to leverage task decomposition and improve reliability without having to develop BARS specific to every task. BARS provide one possible solution for evaluating performance in training in a quantifiable way in order to compare training programs more objectively.

In order to determine how to compare training programs, we finished the section by placing our quantitative solution in the context of Dunne et al.'s (2014) proof of concept study. This study provided an example where quantifiable metrics were used alongside cost avoidance data to prove the return on investing in training programs. We expanded upon this proof of concept study by demonstrating how SPEMS could be used in an overall training system evaluation plan to determine the value of a training system.

We believe that by leveraging the strong inter-rater reliability of military raters demonstrated in Wigdor and Green's study, and the task decomposition / hierarchical approach to training developed by the Marine Corps, we can layer behaviorally anchored rating scales on top of task performance steps to develop a method for measuring task

performance in the military. The quantifiable performance data resulting from this method has the potential to be used similarly to Dunne et al.'s (2014) proof of concept study to determine the performance benefits and avoided costs associated with implementing new training solutions. Discounting the avoided costs based on the relative improvement in performance resulting from training with new systems would link cost data and performance data. Training capability developers could utilize this link to more accurately determine the return on investing in training programs that increase Marine's task proficiency at a reduced cost.

IV. PHASE I: DESIGN OF EXPERIMENTS AND PILOT TESTING

A. OVERVIEW—THE DESIGN OF EXPERIMENTS

This research was conducted in two phases. Phase one consisted of two pilot studies designed to (1) define SPEMS and validate its reliability, and (2) derive the objective measures of performance that are able to measure success in the given task. Phase two consisted of an experiment that was conducted at The School of Infantry West, Camp Pendleton, in order to collect data on the ability of two different performance evaluation systems to capture trainee performance. The experiment manipulated one factor, the performance measurement technique, measured at two levels: current PECL scoring, and the SPEMS. A paired design was used, such that a control evaluator and an experimental evaluator each evaluated the same trainee. A detailed description of the research methodology is discussed in the following sections. All aspects of the research plan were approved by the NPS IRB (NPS IRB#: NPS.2019.0005-IR-EP7-A) and the USMC Human Research Protection Program (HRPP).

B. TWO PILOT STUDIES FOR DEVELOPING BARS RELIABLY

Phase one took place at the Naval Postgraduate School and involved recruiting, consenting, and using a focus group of subject matter experts to develop SPEMS. The focus group conducted one of two tasks—either a card-sorting task or viewing a series of graphically constructed videos. The tasks were conducted iteratively such that the card-sorting task happened first. The results of that task were used for the video review groups, and the results from each video review group were used to conduct each subsequent group. Phase one concluded with a developed, reliable SPEMS and validated measures of performance for use in Phase two.

Phase one consisted of a series of pilot studies in which 17 subject matter experts volunteered to participate in a focus group. The first seven members of the focus group were responsible for developing the behaviorally anchored rating scales as part of the development of SPEMS. As stated in the Chapter II Section E, BARS are typically created through an iterative process that asks participants to create a list of anchors, and then have

a separate group bunch and rank each anchor to assess agreement (Schwab et al., 1975). We utilized a card sorting methodology to complete this developmental process because card sorting is a method for determining how people group and associate specific data on a scale (Usability Professionals Association, 2010). The remaining ten focus group members assisted with assessing inter-rater reliability of SPEMS during Pilot Study 2.

1. Pilot Study 1—Card Sorting: Developing BARS

a. Pilot Study 1—Card Sorting: Methodology

Pilot study 1 consisted of two parts. During the first part, a subset of seven focus group members, split into three teams of two to three members, was asked to card sort behavioral anchors according to their performance level for the development of SPEMS. Each team was asked to develop a list of anchors that could be used to describe the performance of a military task at each level between 1 and 5. Each team was given 30 minutes to write down every anchor they could think of to most accurately capture performance at each level. After a brief pause, each anchor was ranked within each level according to how well the team felt the anchor accurately captured performance at that level. These rankings varied in range depending on how many anchors were listed under each level, with 1 being the anchor which the group felt most accurately captured performance at that level. Focus group members were given a break while the research staff compiled all of the terms together into a new anchor bank consisting of 33 anchors.

For the second part of Pilot Study 1, the same seven focus group members worked individually to sort, group, and rank each anchor of the 33-anchor bank. At this point, individuals were not authorized to add any anchors to their levels, nor could they list any anchors more than once across all levels. Once again, individuals were given 20 minutes to group their anchors into one of the five levels between 1 and 5 according to how well they felt the anchor accurately captured performance at that level. Again, after a brief pause, focus group members were given the opportunity to rank each anchor within each level according to how well they felt the anchor accurately captured performance at that level. Anchors with lower rankings indicated they more accurately captured performance at that level. This concluded Pilot Study 1.

b. Pilot Study 1—Card Sorting: Results

To determine agreement among members of the first focus group, each anchor was analyzed by the level it was placed at and the ranking it received in that level. This analysis was done by counting how many of the seven members placed the same anchor in the same level, and what the mean ranking of each anchor was at each level. Anchors that were placed in the same bin by all seven participants would have 7/7 agreement or 100%. Anchors were retained if they demonstrated a high percentage of agreement (over 50%), and if they received a low mean ranking. Table 3 illustrates the results of the card-sorting task.

Table 3. Card sorting task results—retained anchors

Level	Term	Description	N	% Agreement	Mean Rank	Lower CI	Upper CI
1	A	Performance step not addressed	6	86%	3.5	1.53	5.46
1	E	No acknowledgement	7	100%	4.14	2.11	6.17
1	H	Novice	7	100%	3.5	1.58	5.55
1	T	Unable to execute	7	100%	2	0.58	3.41
2	D	Advanced beginner	7	100%	4.85	1.5	8.2
2	K	Performance step is attempted with a majority of mistakes	6	86%	2	0.37	3.63
2	X	Below standard	6	86%	3.33	1.75	4.91
3	J	Competent	6	86%	1.66	1.13	2.21
3	I	Performance step is attempted with minor mistakes	4	57%	2	0	5.18
4	B	Proficient	7	100%	2.71	1.05	4.37
4	FF	No references required	6	86%	4.16	2.62	5.71
5	C	Performance step is completed with no mistakes	7	100%	2.57	1.27	3.86
5	EE	Flawless Execution	6	86%	2.16	1.13	3.19
5	F	Mastery	7	100%	2.57	0.98	4.16

Participants showed a high degree of agreement across the 14 terms. The mean percent agreement was 91% (s = 12%). The mean ranking was 2.92 (s = 0.98). These results were refined and ordered to develop the initial behavioral anchors shown in Figure 6.

SCORING:
1-Novice: Unable to execute. Performance step not addressed. No acknowledgement.
2-Advanced beginner: Performance step attempted, majority mistakes. Below standard.
3-Competent: Performance step attempted, minor mistakes.
4-Proficient: No references required.
5-Mastery: Flawless Execution. Performance step completed, no mistakes.

Figure 6. Card sorting task results—retained anchors

An interesting result of this card sorting task is that the anchor "No Go" was placed at level 1 by 86% of participants, with a mean rank of 5 (s = 2.28). The anchor "No Go" is currently used to correspond to performance steps that are not performed to standard. "No Go" was omitted from SPEMS due to the potential for historical bias. The establishment of these BARS meant that SPEMS was defined and prepared to be developed through further inter-rater reliability testing.

2. Pilot Study 2: Video Inter-rater Reliability Testing Methodology

a. Creation of the Video Vignettes

Pilot Study 2 consisted of the creation of video vignettes of simulated buddy rushes and three iterations of focus groups using SPEMS to evaluate the simulated buddy rushes. The videos were constructed by training a software development team at NPS, The FutureTech Team at Modeling of Virtual Environments and Simulation (MOVES), on the buddy rush task and captured by simulating buddy rushes in Virtual BattleSpace 3 (VBS3). We developed a series of 15 vignettes designed to illustrate a buddy pair conducting INF-MAN-3001: Conduct fire and movement (buddy rush) at various levels of proficiency. We trained the software development team on what actions would be indicative of different levels of performance for each performance step. The team then played VBS3 to simulate conducting the buddy rush task at various levels of proficiency. The team's actions were captured using standard video playback software to allow the VBS3 simulations to be turned into test videos. A screen shot of one of the videos can be seen in Figure 7. Once the test videos were created, they were further refined to most closely mirror realistic behaviors at various levels of proficiency.

Figure 7. Virtual depiction of "INF-MAN-3001: Conduct Fire and Movement"

b. Assessing SPEMS Inter-rater Reliability

With a set of fifteen fully developed videos, three focus groups of a total of ten infantry officers were convened in order to evaluate each buddy rush using the SPEMS pictured in Figure 8.

INF-MAN-3001: Conduct fire and movement

CONDITION: Given an order from higher and an enemy.

STANDARD: To neutralize the enemy threat in order to accomplish the mission, meeting the commander's intent.

SCORING:
1-Novice: Unable to execute. Performance step not addressed. No acknowledgement.
2-Advanced beginner: Performance step attempted, majority mistakes. Below standard.
3-Competent: Performance step attempted, minor mistakes.
4-Proficient: No references required.
5-Mastery: Flawless Execution. Performance step completed, no mistakes.

PERFORMANCE CHECKLIST (EVENT COMPONENTS)

1. Suppress the enemy (S).	1	2	3	4	5
2. Assess effects of fires (A).	1	2	3	4	5

Note: Performance steps 3–9 omitted due to redundancy.

Figure 8. Iteration 1—SPEMS scoring sheet used by the first focus group

Infantry officers were specifically chosen for these focus groups to leverage the fact that they are trained, baselined, and experienced in evaluating the buddy rush task. The assumption that raters are baselined is an important relevant reality that is referenced by Boldovici et al. (2002), Hubbard (2014), and Wigdor and Green (1991). The three focus groups had a total of 71.5 combined years in the Marine Corps with an average of time in service of seven years, and an average rank of Captain. All of the participants (100%) felt they were qualified to evaluate the task, and on average felt 90% familiar with the task.

The researchers elicited SPEMS scores from each member of the focus group for each video. SPEMS scores were compared and discussed across the focus group to improve the reliability of SPEMS. This focus group process was repeated three times, in which refinement of SPEMS occurred after the first and second iterations in order to improve the inter-rater reliability of SPEMS. The first focus group consisted of four participants, and the subsequent two focus groups consisted of three participants for a total of 10. Each iteration of the focus group concluded with a usability survey about SPEMS and a survey to validate which objective measures of performance measure success in the buddy rush task (see Appendix B). The next section provides a description of the specific procedures used for each focus group, the inter-rater reliability results stemming from that particular focus group, and the changes to the SPEMS evaluation process made based on those results. Finally, survey results regarding SPEMS feasibility, and suggested objective measures of buddy rush performance are described.

Each evaluator was asked to watch each video for an unlimited amount of time and evaluate each performance step of the buddy pair using the SPEMS sheet pictured in Figure 8. The scoring was anchored by the BARS pictured underneath the scoring section of the SPEMS. The total task score was computed by averaging the individual performance step scores for the entire task.

Once all members scored a video, all SPEMS sheets were collected and analyzed for discrepancies of two or more levels on each performance step (i.e., if one member rated a step as a 1 and another member rated it as a 3). If discrepancies were found, participants were asked why they scored tasks the way that they did in order to determine how SPEMS could be refined to improve its inter-rater reliability. After the focus group watched all

fifteen videos, the focus group then completed the SPEMS usability and buddy rush objective measures of performance surveys.

c. Pilot Study 2—Video Focus Groups: Results

We used Cronbach's alpha to assess inter-rater reliability of scoring between focus group members. Descriptive statistics of the survey responses were used to ascertain the ease of use and effectiveness of SPEMS. Finally, descriptive statistics of measure of performance survey responses were used to determine what objective measures of performance should be measured. These results are described iteratively in the following sections; they indicated it was plausible to move on to phase II.

(1) Focus Group 1–4 Participants

(i) Inter-Rater Reliability Results

Inter-rater reliability analysis was conducted by comparing each of the focus group member's overall SPEMS score for each of the 15 videos. The first focus group consisted of four participants who demonstrated an extremely high degree of inter-rater reliability of overall performance scores as evidenced by a Cronbach's Alpha of 0.96. The scatter plot results in Figure 9 indicate that raters in this focus group were evaluating the same underlying concept, the buddy rush, with similar results.

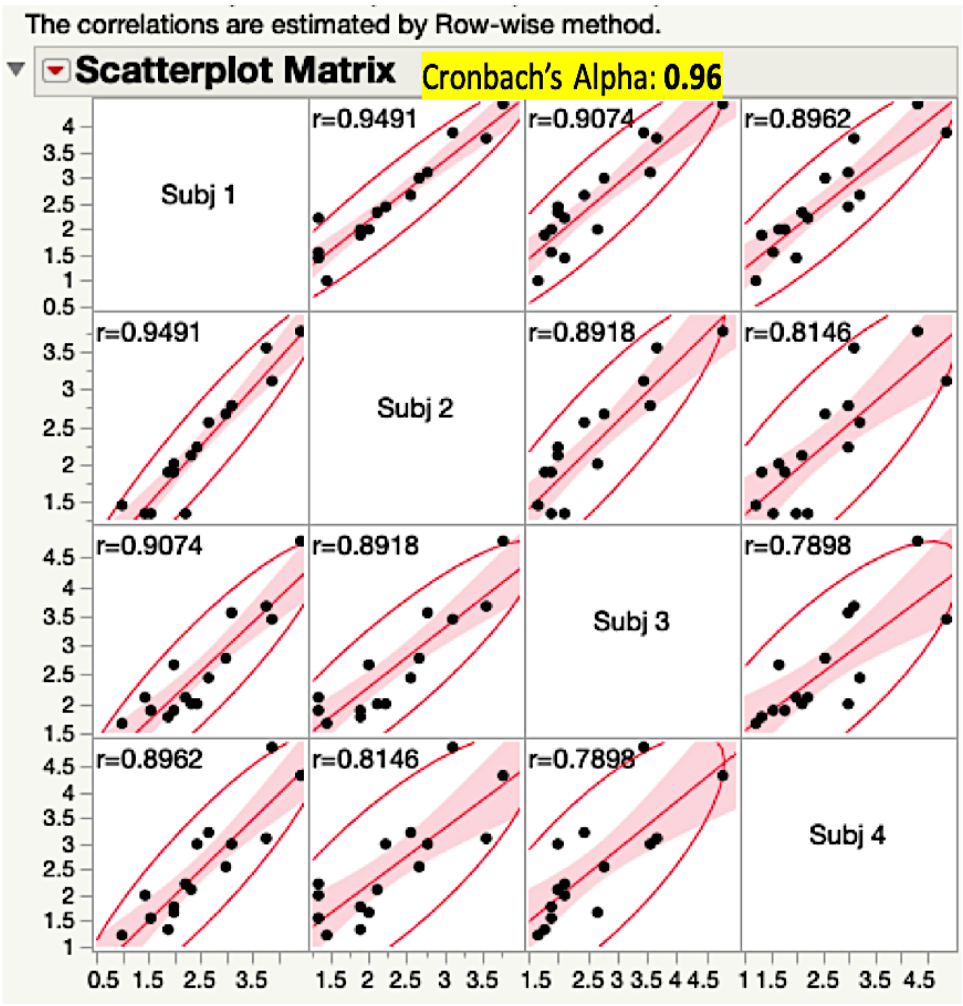
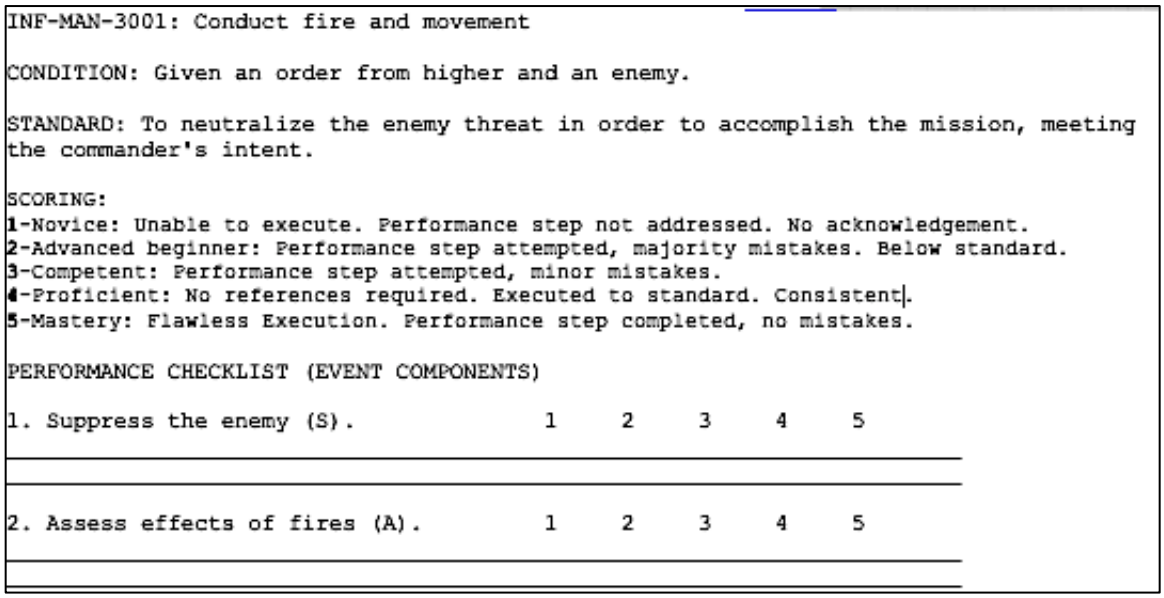


Figure 9. Focus Group 1—Scatterplot results of mean ratings show 0.96 Cronbach Alpha

Feedback from the focus group indicated that the level 4 anchor (4—Proficient: No references required) was not as accurate as the other anchors. While the anchor may have been applicable to other tasks, no references existed as a part of the buddy rush task. This ambiguity made the anchor poorly suited for the video review. Therefore, three changes were made to the procedures: (1) the researchers first provided a review of the task and noted small discrepancies between the virtual range and a real range; (2) the level 4 anchor results of the card sorting task were revisited by the research team in order to provide more anchors to evaluators; and (3) videos with ambiguous context clues were augmented with

verbal injects to mitigate any misrepresented behaviors. These refinements were made for focus group 2, with the SPEMS scoring sheet shown in Figure 10.



INF-MAN-3001: Conduct fire and movement

CONDITION: Given an order from higher and an enemy.

STANDARD: To neutralize the enemy threat in order to accomplish the mission, meeting the commander's intent.

SCORING:
1-Novice: Unable to execute. Performance step not addressed. No acknowledgement.
2-Advanced beginner: Performance step attempted, majority mistakes. Below standard.
3-Competent: Performance step attempted, minor mistakes.
4-Proficient: No references required. Executed to standard. Consistent.
5-Mastery: Flawless Execution. Performance step completed, no mistakes.

PERFORMANCE CHECKLIST (EVENT COMPONENTS)

1. Suppress the enemy (S). 1 2 3 4 5

2. Assess effects of fires (A). 1 2 3 4 5

Note: Performance Steps 3–9 omitted due to redundancy.

Figure 10. Refined SPEMS scoring sheet with new level-4 anchor

(ii) Survey Results

The first focus group's survey responses indicated that SPEMS was an effective performance measurement tool, and that objective measures of performance existed for the task. 100% of participants thought SPEMS was more effective than PECL and on scales ranging from 1 to 10, scored an 8.75 (s = 0.5) for ease of use and a 9.0 (s = 0.0) for effectiveness. Accuracy, time to complete the task, and rate of fire were all chosen unanimously as effective MOPs. Accuracy was ranked the most effective measure (mean ranking = 1.25, s = 0.5), then rate of fire (mean ranking = 3.25, s = 1.5), and then time (mean ranking = 3.5, s = 1.0).

(2) Focus Group 2–3 Participants

(i) Inter-rater Reliability Results

The second focus group consisted of three participants who demonstrated an extremely high degree of inter-rater reliability on overall performance scores as evidenced by a Cronbach's Alpha of 0.93. This reliability was only slightly lower than the first focus group (0.96) and was attributed mostly to the smaller focus group size. The scatter plot results in Figure 11 indicate that raters in this focus group were evaluating the same underlying concept, the buddy rush, with similar results.

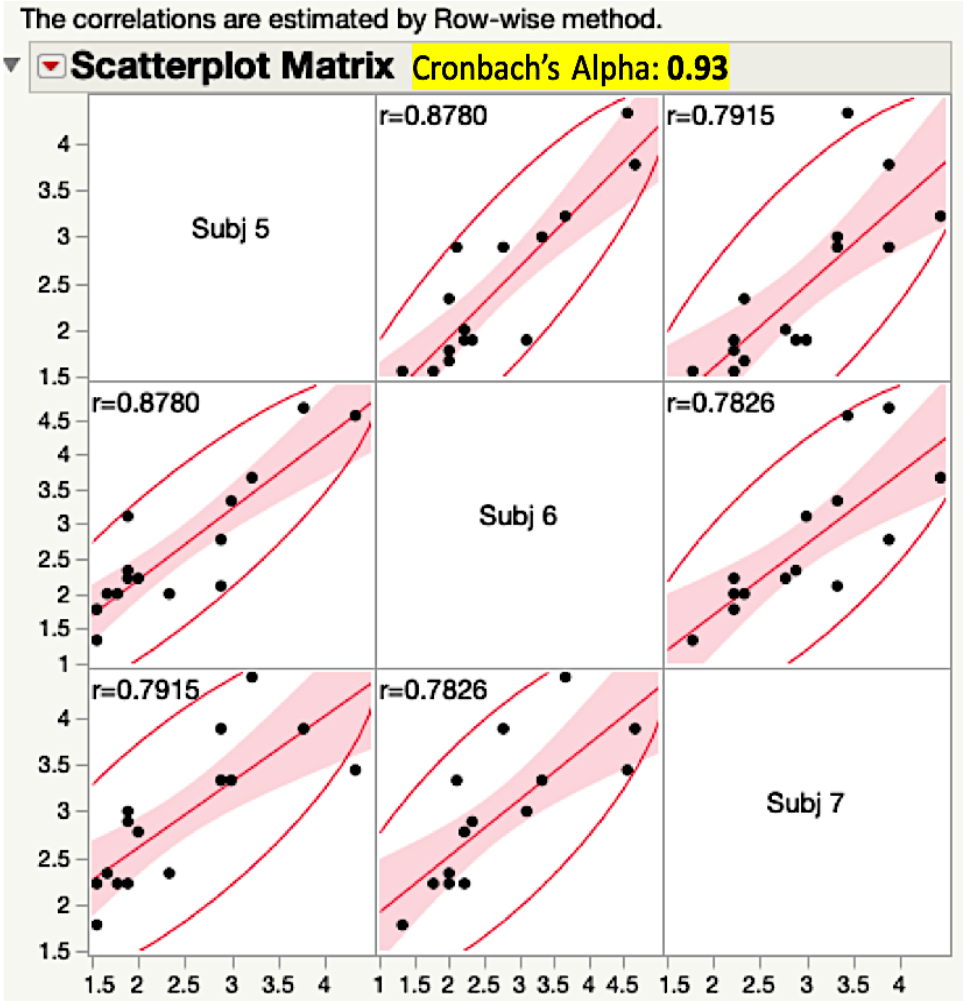
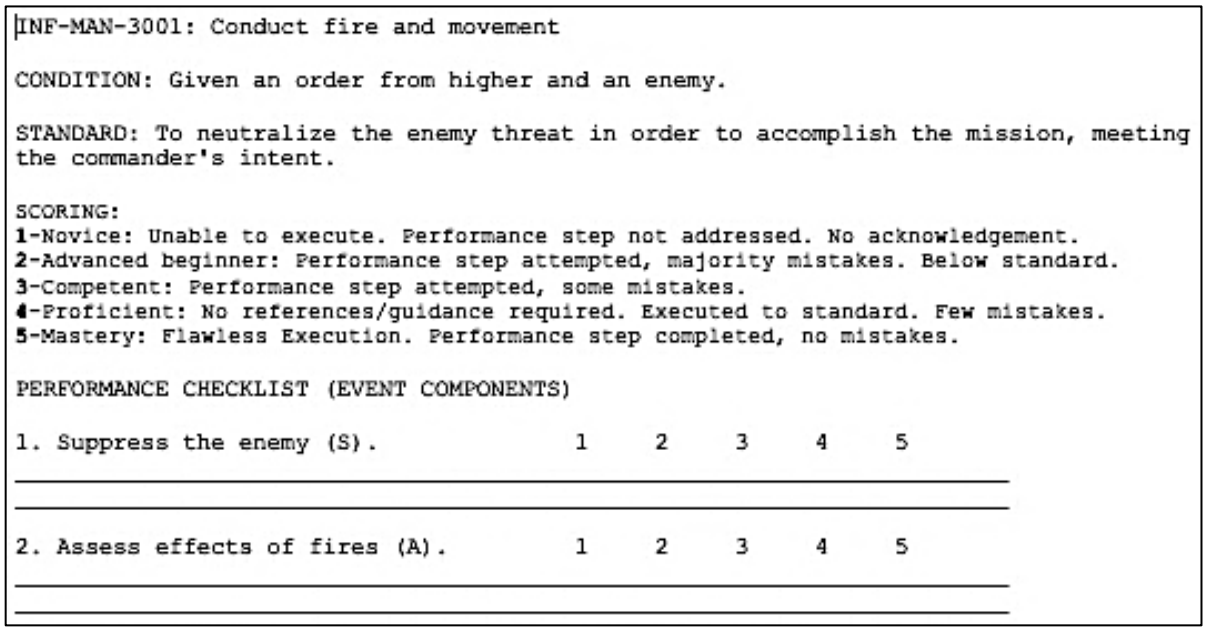


Figure 11. Focus Group 2—Scatterplot results of mean ratings show 0.93 Cronbach Alpha

Feedback from focus group 2 indicated that the level 3 anchor "Minor Mistakes" was misleading evaluators to believe that if one mistake was committed, then the task must be evaluated at a 3. Further, the level 4 anchor was refined in order to address feedback that the anchor "No references required" was not applicable to this task. Finally, both focus groups 1 and 2 had demonstrated a high degree of inter-rater reliability and consistent scoring for the first five videos. As a result, we chose to use these videos as training to baseline the final focus group in an effort to further improve reliability. With these refinements in mind, we proceeded to focus group 3 with the SPEMS scoring sheet, shown in Figure 12.



INF-MAN-3001: Conduct fire and movement

CONDITION: Given an order from higher and an enemy.

STANDARD: To neutralize the enemy threat in order to accomplish the mission, meeting the commander's intent.

SCORING:
1-Novice: Unable to execute. Performance step not addressed. No acknowledgement.
2-Advanced beginner: Performance step attempted, majority mistakes. Below standard.
3-Competent: Performance step attempted, some mistakes.
4-Proficient: No references/guidance required. Executed to standard. Few mistakes.
5-Mastery: Flawless Execution. Performance step completed, no mistakes.

PERFORMANCE CHECKLIST (EVENT COMPONENTS)

1. Suppress the enem (S). 1 2 3 4 5

2. Assess effects of fires (A). 1 2 3 4 5

Note: Performance Steps 3–9 omitted due to redundancy.

Figure 12. Refined SPEMS scoring sheet with refined level 3 and 4 anchor.

(ii) Survey Results

The second focus group's survey responses indicated that SPEMS was an effective performance measurement tool, and that objective measures of performance existed for the task. 100% of participants thought SPEMS was more effective than PECL and scored an 8.66 (s = 1.15) for ease of use and a 9.33 (s = 1.15) for effectiveness. Accuracy, rate of fire,

and the number of communication events were all chosen unanimously as effective MOPs. Accuracy was ranked the most effective measure (mean ranking = 1.33, s = 0.58), then rate of fire (mean ranking = 2.66, s = 0.58), and then communication (mean ranking = 3.66, s = 0.58).

The research team discussed the feasibility of measuring the MOPs emerging from the survey results and decided to add some other measures of performance to the survey. Time to complete the task would be uniform across all buddy pairs because they completed the task in a uniform amount of time as their squad. Rate of fire would be difficult to measure given the research team's limitations in audibly measuring bullets fired by a specific buddy pair per minute. Furthermore, this measure would change throughout the completion of the task. Finally, accuracy was the only measure left; however, based on the limitations of the range, accuracy could only be measured by the number of times the target bobs. In order to determine what to measure more effectively, qualitative survey feedback was reviewed for additionally suggested measures. Two subjects suggested measuring the time a buddy pair is moving with the target up, and the total number of rushes. These measures were added to the survey, and time and communication events were removed because they could not be accurately measured or were uniform for all buddy pairs.

(3) Focus Group 3–3 Participants

1. Inter-rater Reliability Results

The third focus group consisted of three participants who demonstrated an extremely high degree of inter-rater reliability on overall performance scores as evidenced by a Cronbach's Alpha of 0.98. This reliability was higher than the first and second focus group (0.96 & 0.93 respectively) which is attributed mostly to the fact that the first 5 videos were used to train the final focus group, and the results were based on their scoring of the last 10 videos. The scatter plot results in Figure 13 indicate that raters in this focus group were evaluating the same underlying concept, the buddy rush, with similar results. Due to the almost perfect indications of inter-rater reliability, no changes were made to SPEMS as a result of focus group 3.

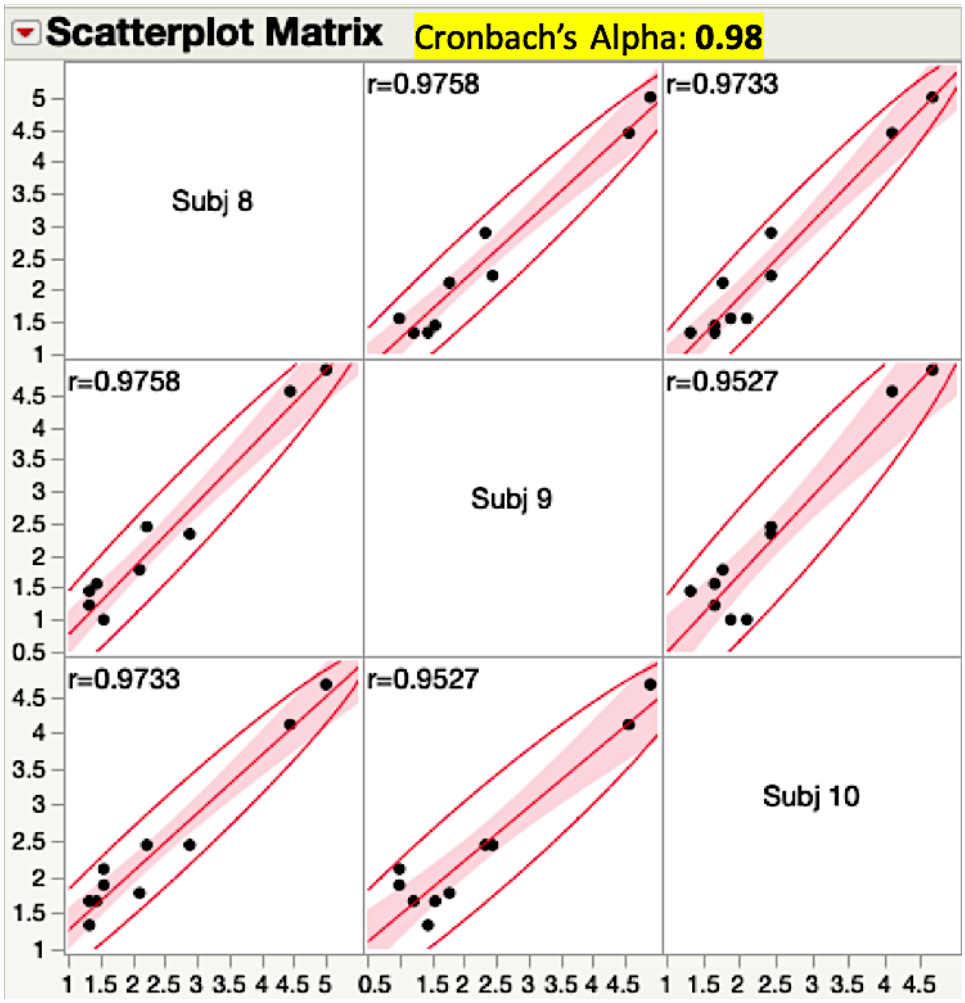


Figure 13. Focus Group 3—Scatterplot results of mean ratings show 0.98 Cronbach Alpha

The final focus group indicated such a high degree of inter-rater reliability that no changes were made to SPEMS prior to phase 2. Showing videos to baseline evaluators proved to be a successful method for improving inter-rater reliability indicating that videos should be used to baseline SPEMS evaluators prior to phase 2. The purpose of phase 2 is to test whether SPEMS accurately evaluates live-execution of the buddy rush task in the operational environment. This accuracy was tested by measuring SPEMS' relationship with agreed upon objective measures of performance.

2. Survey Results

The third focus group's survey responses indicated that SPEMS was an effective performance measurement tool, and that objective measures of performance existed for the task. 100% of participants thought SPEMS was more effective than the PECL and scored it a 9.0 (s = 1.0) for ease of use and a 9.0 (s = 1.0) for effectiveness. Accuracy, number of rushes, and time spent moving with the target up were all chosen unanimously as effective MOPs. Time spent moving with the target up was ranked the most effective measure (mean ranking = 1.0, s = 0.0), then accuracy (mean ranking = 2.66, s = 1.15), and then number of rushes (mean ranking = 6.0, s = 0.0). Based on previous feedback, researcher's ability to measure, and the unanimous selection of time spent moving with the target, time spent moving with the target up was chosen as the primary MOP for the buddy rush task.

C. PILOT STUDY RESULTS OVERVIEW

1. SPEMS Development and Inter-rater Reliability

SPEMS development began with Pilot Study 1: Card Sorting and was refined through Pilot Study 2: Video Focus Groups. For Pilot Study 1, the mean percent agreement for sorting anchors was 91% (s = 12%), and the mean ranking for each anchor was 2.92 (s = 0.98). This high degree of agreement allowed researchers to confidently select anchors that were refined during the second pilot study.

For Pilot Study 2, inter-rater reliability was measured by calculating the Cronbach's alpha for each focus group. The first focus group consisted of four participants who demonstrated an extremely high degree of inter-rater reliability of overall performance scores as evidenced by a Cronbach's Alpha of 0.96. The second focus group consisted of three participants who demonstrated an extremely high degree of inter-rater reliability on overall performance scores as evidenced by a Cronbach's Alpha of 0.93. The third focus group consisted of three participants who demonstrated an extremely high degree of inter-rater reliability on overall performance scores as evidenced by a Cronbach's Alpha of 0.98. This high degree of inter-rater reliability across all focus groups validated the reliability of SPEMS prior to operational experimentation.

2. SPEMS Ease of Use and Effectiveness

100% of participants across all focus groups indicated that they felt SPEMS was more effective than the PECL evaluation method. On a scale between 1 and 10, SPEMS was scored with a mean of 8.8 ($s = 0.76$, 95% CI = (8.33, 9.26)) for ease of use, and a 9.1 for effectiveness ($s = 0.69$, 95% CI = (8.67, 9.53)) at evaluating training. Finally, focus groups felt that, on average, evaluators could be reliably trained to use SPEMS in under one hour using the videos.

3. Buddy Rush Measures of Performance

Survey data from focus group 3 indicated that the measure of performance most indicative of buddy rush task performance was the amount of time a buddy pair rushes and is exposed to the target. 100% of the final focus group agreed that the amount of time a buddy pair is exposed measures performance of the task, and the amount of time a buddy pair is exposed was ranked the most indicative measure of performance of the six measures to choose from. Due to the inability of researchers to accurately add up the amount of time a buddy pair conducts an exposed rush, the percentage of rushes that were exposed was used as a proxy since the time it takes to conduct a buddy rush is controlled by the training environment. This proxy measurement was validated by further soliciting subject matter expert feedback.

The percentage a buddy pair is exposed is measured by calculating the proportion of total exposed rushes over the total number of rushes. An exposed rush is defined as one buddy advancing towards the target without suppression. “Without suppression” is determined by observing if the pop-up target is up at the same time that a buddy is moving. If a target is up while one of the Marines in the buddy pair is moving, the rush is counted as an exposed rush. If a Marine who is behind cover shoots the target causing it to go down, and the other Marine advances to a piece of cover prior to the target coming back up, the rush is not counted as an exposed rush. Therefore, the percent of exposed rushes was the main measure of performance for Phase 2, in which lower percentages indicate more effective buddy rush performance.

THIS PAGE INTENTIONALLY LEFT BLANK

V. PHASE II: LIVE EXPERIMENTATION AND RESULTS

A. INTRODUCTION

Phase two consisted of the experiment conducted at the School of Infantry (SOI) West. Evaluators evaluated trainees' ability to conduct established training procedures to train and qualify in the task: INF-MAN-3001. There was a total of four evaluators randomly assigned to two groups–two SPEMS evaluators, and two PECL evaluators. For each run, a PECL evaluator and a SPEMS evaluator evaluated one buddy pair of two trainees conducting the task. A total of 26 buddy pairs (52 Marines) were evaluated three times. The first evaluation was conducted during the trainees' final blank-fire or practice run. The remaining two evaluations were conducted under live-fire conditions.

The experiment was conducted under the following conditions and with the following population. No modifications were made to the current training curriculum outside of increased scoring methodologies and evaluators. Evaluators were given the same amount of time that they are currently given to evaluate the task.

B. EXPERIMENT PROCEDURES AND REMARKS

1. Participants

There were two sets of participants, evaluators and trainees. Evaluators were all the rank of Sergeant (E-5) with approximately five to seven years of experience. Trainees were all 0311 infantryman who had recently completed boot camp and were completing their 0311 MOS training to join the operational forces. Data was collected on 52 trainees paired together into 26 buddy pairs. Trainees were all blind to the scoring methods and the scoring in general to further reduce the impact of the study on the training environment. These conditions were maintained during the following empirical process.

2. Procedures

The experiment began with recruitment, the consent process, population verification, and training. Trainees were recruited by gathering all of the third fireteams of the training company. One fireteam per squad was specifically selected to most efficiently

spread researchers and gather the most data. Trainees were read a recruitment script explaining they would conduct their standard training procedures, and that they were consenting to having data gathered on them. The trainees were then provided the opportunity to consent to being evaluated using SPEMS / PECL and having their performance measured according to the MOPs. The trainees were required to perform no tasks outside of their standard training besides consenting. The School of Infantry West, Infantry Training Battalion, asked for combat instructors from their instructor pool to volunteer to participate in the study. Once the research team met the volunteers, they were read their recruitment script and given the opportunity to consent to participating in the study. All consented to participate in the study. Evaluators were randomly assigned to two groups, PECL evaluators and SPEMS evaluators. Evaluators were trained using the following procedures.

PECL evaluators verbally verified that they were trained in evaluating the task using a PECL and were dismissed in order to maintain their blindness to SPEMS. SPEMS evaluators were trained on how to use SPEMS using the validated video data from pilot study 2. SPEMS evaluators were shown five videos and told the mean SPEMS score each video was given by the focus groups. The videos were approximately spread across each SPEMS grading level one to five in order to demonstrate to the SPEMS evaluator what performance of the buddy rush task looked like at each evaluation level. SPEMS evaluators were given the opportunity to score each video to practice using the scale. The training took approximately 30 minutes at which time the SPEMS evaluators were dismissed in order to maintain their blindness to what objective measures of performance were being collected. Finally, the research team practiced counting a buddy rush and counting an exposed buddy rush. This practice consisted of watching trainees conduct buddy rushing with the research team to ensure counting was happening uniformly for both lanes. This practice was done in the absence of all evaluators to ensure evaluators were not influenced in their scoring by knowing that percentage exposure was being measured.

The 0311-training company consisted of thirteen squads each consisting of 3 fireteams. A fireteam consists of two buddy pairs or four people. The third fireteam of each squad was randomly selected to be evaluated for a total of 26 buddy pairs or 52 trainees to

be evaluated. The 26 buddy pairs were split between two lanes of 13 buddy pairs per lane as a part of the standard buddy rush training. Each lane was assigned a PECL evaluator, a SPEMS evaluator, a number of rushes counter, and a number of exposed rushes counter. The 26 buddy pairs were evaluated three times, once during their final blank fire practice run, and twice during live-fire conditions. During each live-run, the data collectors counted the total number of rushes per buddy pair, and the total number of exposed rushes per buddy pair. The percentage exposure was then calculated using the following equation. $\%\, Exposure = \frac{Number\ of\ Exposed\ Rushes}{Total\ Number\ of\ Rushes}$. The experiment was completed after the last buddy pair was evaluated and measured on the third run. After the experiment, the two SPEMS evaluators were provided a survey to give feedback on the usability of SPEMS.

3. Statistical Methods and Assumptions and Conditions

The paired t-test and linear regression were the primary statistical methods used in the following results section. All statistical methods were tested using two tailed tests at an alpha level 0.05. The assumptions and conditions for a paired t-test are: paired continuous data, sample size larger than 15, and normality. The assumptions and conditions for linear regression are linearity, independence, normality, and equal variance. These assumptions and conditions were checked, discussed, and the results are shown in Appendix D.

C. PRELIMINARY RESULTS

1. SPEMS and PECL Results

The distribution of scores overall and by performance step by evaluation type (SPEMS or PECL) are described in Table 4. The PECL score was calculated by treating every "Go" as a 1 and every "No Go" as a 0 and averaging the score across the performance steps. Descriptive statistics (means and standard deviations), as well as paired-*t* tests were used to determine statistical significance. The Shapiro-Wilk goodness of fit test for normality indicated that the difference between runs one and two for PECL did not meet normality and for SPEMS approached non-normality. Therefore, a Wilcoxon signed rank test was used to demonstrate the difference between runs one and two. The assumptions and conditions were met for all other t-tests.

Table 4. Performance step evaluation descriptive statistics by evaluation method and run

Run	(N = 26)	Performance Step 1	PS 2	PS 3	PS 4	PS 5	PS 6	PS 7	PS 8	PS 9	Overall
1	PECL Mean	0.92	0.50	0.27	0.92	0.65	0.62	0.69	0.62	0.88	0.68
	PECL SD	0.27	0.51	0.45	0.27	0.48	0.50	0.47	0.50	0.33	0.24
	SPEMS Mean	3.15	3.15	3.19	3.42	2.81	2.73	3.15	3.35	3.35	3.15
	SPEMS SD	0.88	0.612	0.633	0.58	0.69	0.67	0.61	0.56	0.49	0.42
2	PECL Mean	0.88	0.81	0.81	1	0.85	0.85	0.54	1	1	0.86
	PECL SD	0.33	0.40	0.40	0	0.37	0.37	0.51	0	0	0.18
	SPEMS Mean	3.42	3.50	3.62	4.12	3.50	3.46	3.62	4.04	4.15	3.71
	SPEMS SD	0.90	1.10	1.02	0.86	0.86	0.86	0.64	0.53	0.67	0.53
3	PECL Mean	0.85	0.85	0.88	0.92	0.61	0.77	0.38	1	1	0.81
	PECL SD	0.37	0.37	0.33	0.27	0.50	0.43	0.50	0	0	0.18
	SPEMS Mean	3.31	3.31	3.27	4.12	3.00	3.00	3.15	3.88	4.50	3.50
	SPEMS SD	0.84	0.68	0.78	0.95	0.63	0.63	0.78	0.82	0.71	0.43

Overall, the evaluators rated the trainee's performance as improved from run 1 to run 2 (Wilcoxon(SPEMS) S = 123.5, p = 0.0006; Wilcoxon(PECL) S = 118, p = 0.001). There was no significant improvement in SPEMS or PECL overall scores from run 2 to run 3 (t-SPEMS(25) = 1.51, p = 0.146; t-PECL(25) = 0.81, p = 0.43). It should be noted that run 1 was the blank fire practice run which made it more difficult for evaluators to evaluate performance.

SPEMS ratings showed sensitivity at the performance step level. Performance step 9, "Consolidate" received the highest SPEMS mean score of 4.15 in run 2 and 4.50 in run 3. Performance step 6, "Identify your target and continue suppression in order to allow your buddy to move," received the lowest SPEMS mean score of 3.46 in run 2 and 3.00 in run 3. It is interesting to note that every single buddy pair in the sample received a "Go" from the PECL evaluators for performance step 8 and 9, "Conduct actions on the objective (K)" and "Consolidate" in both runs 2 and 3. This finding could suggest that evaluators did not have enough evidence to rate a single buddy pair as "No Go," and therefore rated them all as "Go" regardless of the potential differences in their performance.

The distribution of mean PECL scores and mean SPEMS scores by run are shown side-by-side in Figure 14. Descriptive statistics reveal much more variability in the PECL rating than the SPEMS rating. PECL standard deviations are about 21% of the mean vs SPEMS at about 13%.

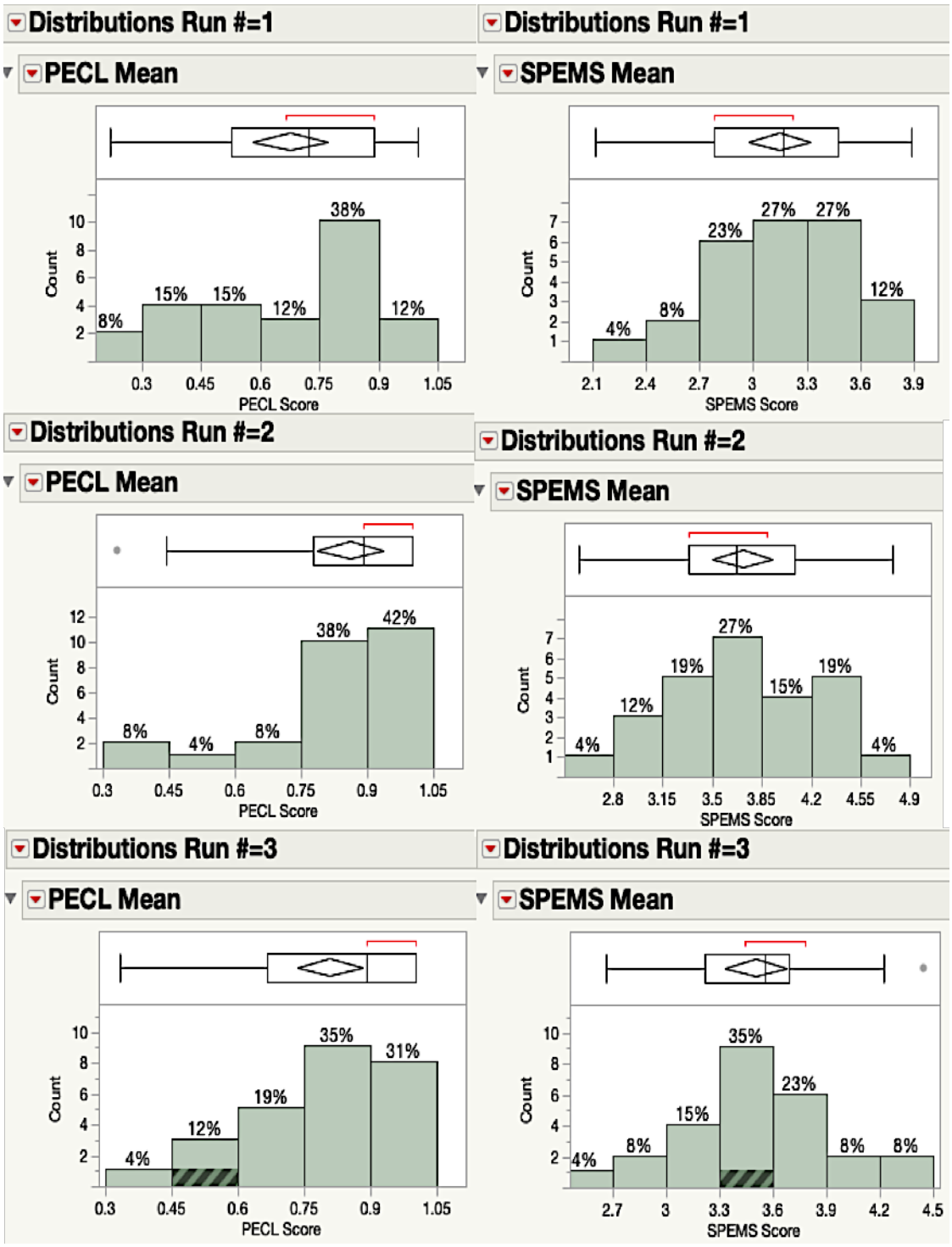


Figure 14. SPEMS and PECL distribution by run indicating an approximately left skewed distribution for PECL as compared to the approximately normal distribution for SPEMS

The key finding to note when comparing the distributions and descriptive statistics between PECL scores and SPEMS scores is the difference in the distributions. It is assumed that trainee performance would follow a normal distribution similarly to Figure 15.

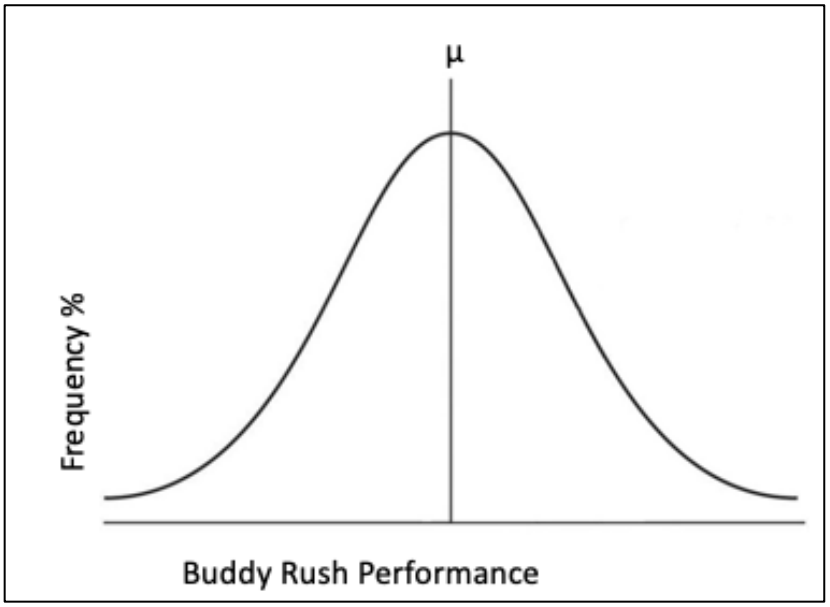


Figure 15. Estimated trainee performance distribution

Instead, PECL scores are left-skewed indicating a large percentage of the distribution is in the highest scoring bin (An average of 30% (s = 15.4) of the buddy pairs received a perfect score of 1.0). This skewness could indicate a tendency for PECL evaluators to provide more "Go" evaluations then trainee performance should warrant. Conversely, SPEMS scores show an approximately normal distribution with the largest percentage of the distribution being centered at a 3.5 rating. This approximately normal distribution suggests that SPEMS more accurately captures buddy rush trainee performance than the PECL.

2. Percent Exposure Results

The distribution of the number of rushes and the number of exposed rushes, is described Table 5 which shows the descriptive statistics for each run and each measure. Note that there are no measures of performance for run 1 because run 1 was the final blank

fire/ practice run. As a result, exposed rushes could not be counted because no rounds were being fired and the target would not go down. SPEMS and PECL data was gathered for run 1 in order to test both metrics ability to predict future live-fire performance from non-live fire practice.

Table 5. Measures of performance descriptive statistics by measure and run

Run	(N=26)	
2	# of Rushes Mean	10.69
	# of Rushes SD	1.76
	# of Exposed Rushes Mean	6.54
	# of Exposed Rushes SD	1.63
3	# of Rushes Mean	10.73
	# of Rushes SD	1.22
	# of Exposed Rushes Mean	5.81
	# of Exposed Rushes SD	2.04

These descriptive statistics begin to shed some light on what average performance looks like. There was no statistically significant difference in the average percent exposure between Run 2 (61%) and Run 3 (54%) ($t(25) = 1.71$, $p = 0.09$). Again, we assume that percent exposure follows a normal distribution. Figures 16 and 17 provide histograms and descriptive statistics for percent exposure for runs 2 and 3.

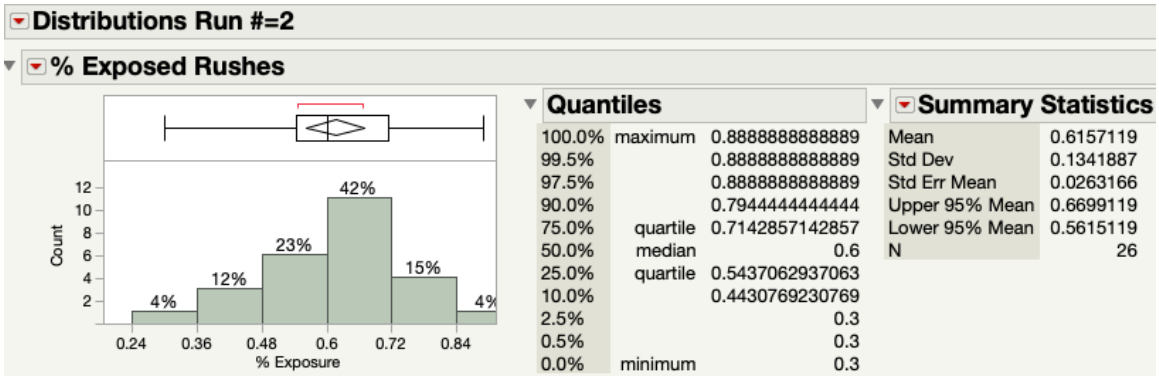


Figure 16. Percent exposed rushes distribution for Run 2 and descriptive statistics

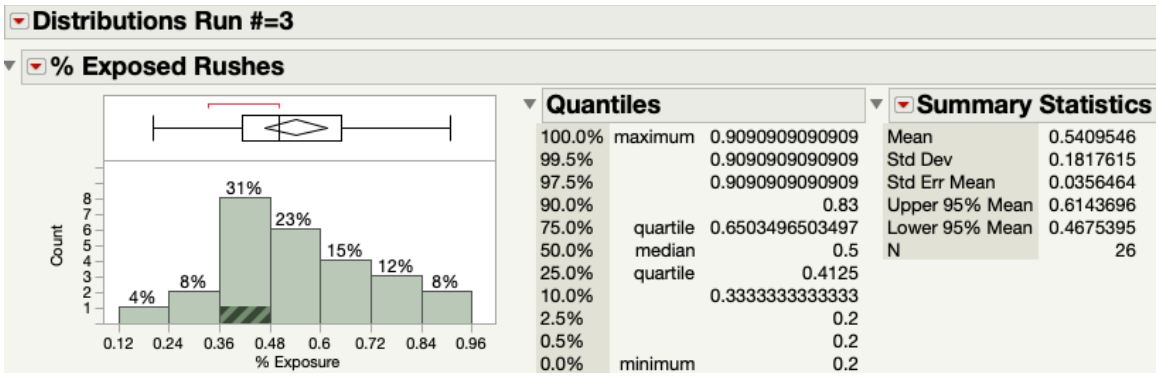


Figure 17. Percent exposed rushes distribution for Run 3 and descriptive statistics

These distributions both visually appear to be approximately normal. If we compare these distributions to the PECL and SPEMS score distributions shown in Figure 14, SPEMS' normal distribution most closely resembles the percent exposure normal distribution. This finding would suggest that there may be a relationship between SPEMS and percent exposure. Conversely, the PECL distribution is left skewed. We conclude the preliminary findings to test our primary two hypotheses.

D. RESULTS

The primary results are centered around testing the original hypothesis laid out in chapter 1 of this thesis. Restated, the hypotheses are:

1. Hypothesis Testing

a. Hypothesis 1

- H_0: There is no relationship between SPEMS scores and objective measures of performance.
- H_A: There is a relationship between SPEMS scores and objective measures of performance.

b. Hypothesis 2

- H_0: There is no difference in the predictive strength between SPEMS scores and PECL on objective measures of performance.
- H_A: There is a difference in the predictive strength between SPEMS scores and PECL on objective measures of performance.

2. Hypothesis 1 Results

We tested this hypothesis using a linear regression model. The linear regression models that were used are shown in the following equations:

Run 2: % Exposure = 1.216–0.162*SPEMS Score

Run 3: % Exposure = 1.480–0.268*SPEMS Score

As previously stated, the assumptions and conditions for linear regression were checked, met, and the detailed analysis is shown in Appendix D. The results from the linear regression models are shown in Figure 18.

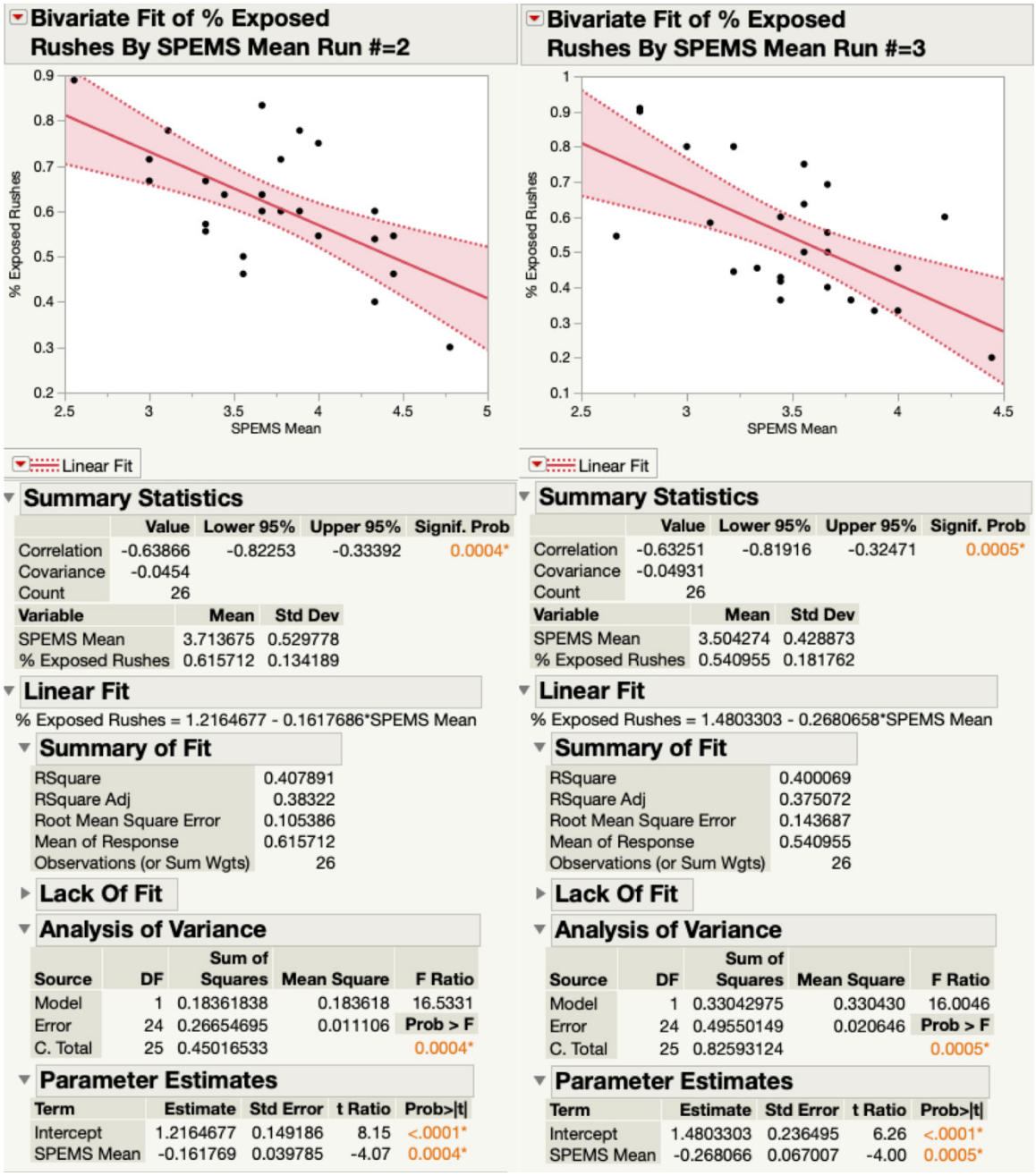


Figure 18. Linear regression results testing fit of mean SPEMS score to percent exposure. Both runs show R^2=0.40 demonstrating good fit.

Run 2 and run 3 have an R^2 of 0.41 and 0.40 respectively demonstrating a moderate, negative, linear relationship between SPEMS scores and percent exposure: as SPEMS scores increase, the percent exposure during the conduct of a buddy rush decreases. For example, in run 3, the model predicts that with each additional point of SPEMS, there is a

26.8% decrease in percent exposure. The R-square values indicate that SPEMS accounts for approximately 40% of the variability in percent exposed rushes. Finally, results of the t-test of the slope indicate a statistically significant negative relationship between SPEMS scores and percent exposure (Run 2: $t(24) = -4.07$, $p = .0004$; Run 3: $t(24) = -4.00$, $p = .0005$). *Therefore, we reject our null hypothesis and conclude that there is a negative relationship between SPEMS scores and percent exposure.*

3. Hypothesis 2 Results

To test hypothesis 2, we conducted linear regressions between PECL scores and percent exposure and compared the results to those using SPEMS scores as the predictor variable. Again, we first checked that this data met the assumptions and conditions for linear regression. For Run 2, the PECL data did not adequately meet the linearity, equal variance, or independence assumption and therefore, linear regression should not be employed (see Appendix D). For Run 3, there are concerns regarding the equal variance and independence assumptions. Therefore, the Run 3 PECL linear regression results should be interpreted with caution. The PECL linear regression model for Run 3 is shown in the following equation and the results are described in Figure 19.

Run 3: % Exposure = 0.901–0.446*PECL Score

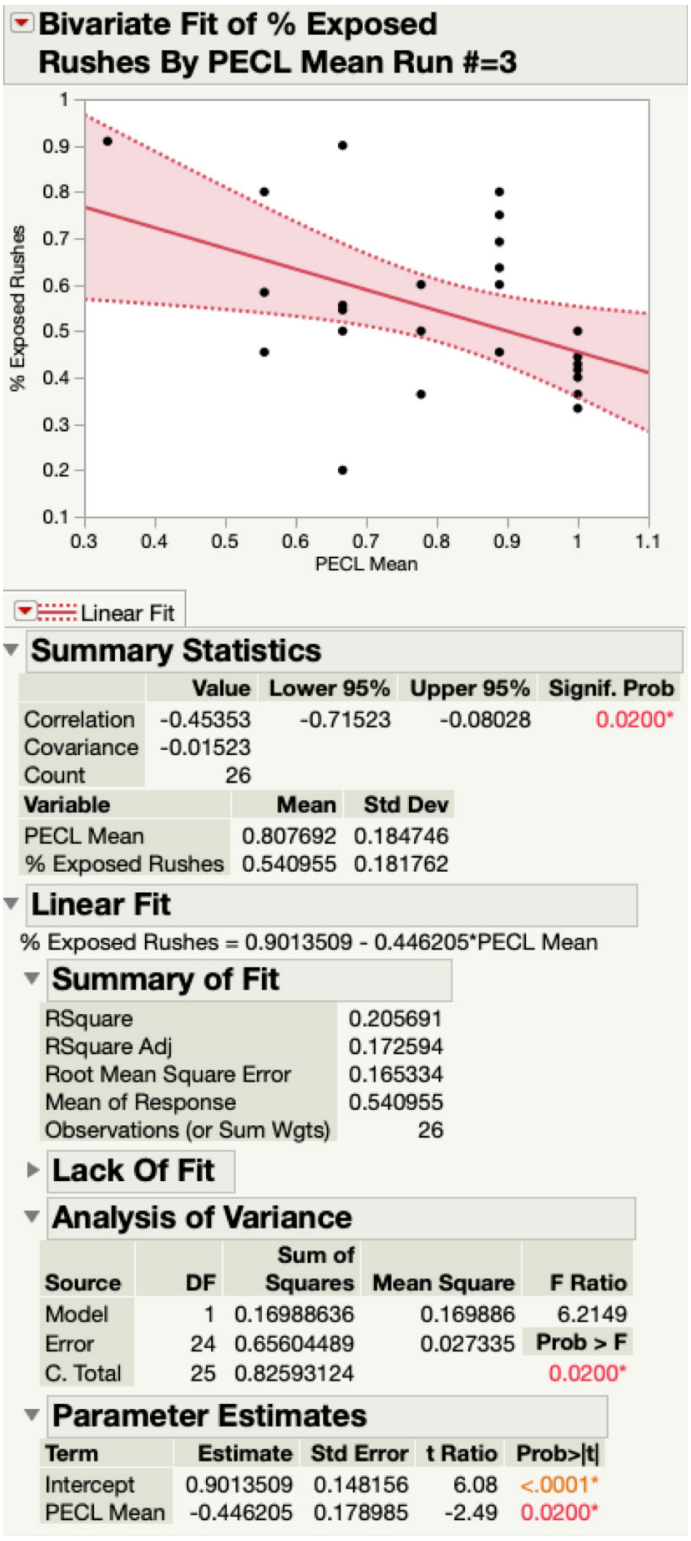


Figure 19. Linear regression results testing fit of mean PECL score to percent exposure. Run 2 was rejected and run 3 shows R^2=0.2 demonstrating poor or no fit.

Run 3 has an R^2 of 0.21 demonstrating a weak, negative, linear relationship between PECL scores and percent exposure. As PECL scores increase, the percent exposure during the conduct of a buddy rush stays approximately the same or decreases slightly. For example, the model predicts that with each additional PECL point, there is a 44.6% decrease in percent exposure. The R^2 values indicate that PECL scores only explain 21% of the variability in percent exposure. Slope results in Figure 19 show a test statistic of -2.49 (p =.020). Because the PECL data does not adequately meet the assumptions and conditions, this statistical result should be viewed with caution. If we revisit our final hypothesis and compare these results in Figure 19 to the linear regression of SPEMS scores shown in Figure 18, we demonstrate that SPEMS has more predictive power than the PECL. *We reject the null hypothesis and conclude that SPEMS scores have more predictive strength then PECL scores.*

THIS PAGE INTENTIONALLY LEFT BLANK

VI. DISCUSSION, RECOMMENDATIONS, AND FUTURE WORK

This thesis research demonstrated a verifiable, repeatable, and reliable method for measuring military task performance across training solutions. BARS were developed for SPEMS using card-sorting techniques; SPEMS' reliability was validated through virtual video analysis; and SPEMS' predictive strength was empirically proven through operational testing. The introduction, problem statement, and background chapters described the scope of the research and justified the research staff's method for developing SPEMS. The methodology and results chapters cover the empirical application of SPEMS to prove it is a consistent and reliable evaluation tool that has more predictive strength than current performance evaluation methods. The following sections discuss the details surrounding the development of SPEMS while providing organizational recommendations for future work in this area of research.

A. DISCUSSION

1. Pilot Testing

a. Focus Groups' Implementation of "No Go" as an Anchor and Its implications

The card-sorting task that defined the BARS for SPEMS provided some useful and interesting findings. The anchor "No Go" was not initially included in the provided bank of words but was added by the participants prior to the second iteration. During the second iteration, 86% of participants placed the anchor "No Go" in level 1 with a mean rank of 5.00 (s = 2.28). This finding would indicate that "No Go" is commonly agreed to correspond to a level 1 out of 5 or 20%. Therefore, achieving a "Go" on a PECL would correspond to achieving any score higher than a 1 out of 5 or a performance step being completed with more than 20% proficiency. *The Marine Corps anecdotally utilizes 80% as its passing standard, and this finding would indicate a difference in what is considered passing by sixty percentage points.* Further testing in the operational environment would be required to prove this difference. "No Go" was omitted from SPEMS due to historical

bias. All other anchors were retained based on agreement and ranking for the definition of SPEMS.

b. Assumption of Inter-rater Reliability

The second pilot test used virtual video analysis to determine the inter-rater reliability of SPEMS as well as surveys to garner feedback from SMEs on SPEMS and buddy rush MOPs. A disagreement between raters by two or more levels on a given performance step initiated a focus group discussion. Focus groups were asked to concentrate on how SPEMS itself could be modified to provide clarity and assist in more accurately evaluating performance and ignore minor video errors. This feedback along with SPEMS survey responses triggered the aforementioned SPEMS' modifications (see Chapter IV, Section B) as well as called for an initial discussion to further baseline focus group participants. This request was based on the potential for participants' familiarity with the task to have decayed while at the Naval Postgraduate School. With the criticality of baselining in mind, research staff decided to use reliably evaluated videos to baseline future evaluators.

After the second focus group results indicated a consistently high level of inter-rater agreement, the first five videos were used as training. The final focus group participants were told the average rating for the first five videos in order to further baseline the raters prior to the final 10 videos. This adaptation led to the highest levels of inter-rater agreement and the determination that SPEMS was a reliable performance evaluation tool. This result further supported one of SPEMS' critical underlying assumptions. *The reliability of the SPEMS is dependent on the standardization and baselining of the evaluators. Infantry officers were specifically chosen as the evaluators to ensure that evaluators were trained, baselined, and qualified to evaluate the task. The operational forces only use trained evaluators making this a valid assumption.* To test the accuracy of SPEMS, scores would have to be compared to an objective measure of buddy rush performance to see if a relationship existed.

c. Feasibility Issues in Determining a Buddy Rush Measure of Performance

SME survey results originally indicated that accuracy was best measure of buddy rush performance. The research team determined that accuracy could not be easily measured in the current training environment without causing interruptions. SME comments further indicated that the time a buddy pair spends rushing while also being exposed might be a more plausible measure. The final focus group's survey included this measure, and 100% of the focus group stated that exposure was the most accurate measure of buddy rush performance. The amount of time exposed was modified to percentage exposure based on the research staff's limited ability to assess the exposure and add up all of the time simultaneously. Percentage exposure was selected as the objective MOP for the buddy rush task based on SME feedback.

Based on the SMEs' combined 71.5 years of infantry experience, we accept that percentage exposure can serve as an objective proxy measurement for performance of the buddy rush task. As a result, as the percentage a buddy pair is exposed increases, the level of performance decreases. We demonstrated a moderate, negative, linear relationship between SPEMS scores and the percent a buddy pair is exposed. Therefore, assuming SME feedback is correct, there is a positive relationship between SPEMS scores and the performance of the buddy rush task. As the level of performance of the task increases, SPEMS scores should also increase. Showing this relationship exists empirically demonstrates that SPEMS accomplished the measurement of buddy rush task performance as was intended.

2. Experiment—Operational Testing

The experiment leveraged existing training processes at SOI West to test the accuracy and usability of SPEMS. Integrating SPEMS into the operational environment allowed researchers to gain insights into how improvements on the individual performance step level influenced overall task performance. For example, performance step 9, "Consolidate" and performance step 6 "Identify next covered position" saw the greatest improvement from run 1 to run 2. In contrast, performance step 1, "Suppress the enemy"

and “Assess the effects of fires” saw the least improvement. This targeted performance data could allow the training staff to focus future deliberated practice prior to live fire qualification on these highlighted areas.

Conversely, by integrating the experiment within the operational environment real world limitations were placed on what MOPs could be measured. Percentage exposure was selected by the staff based on being the SME selected MOP that could be accurately measured in the operational environment. To measure percentage exposure, the research staff had to count each rush and each exposed rush. Counting an exposed rush is dependent on the target correctly responding to accurate rounds by going down. Otherwise, the target may be up indicating there is not suppression when there actually is.

The moderate strength of the relationship between SPEMS scores and percent exposure evidenced by an $R^2 = 0.41/0.40$ may have been weakened by error associated with the targets. Targets sense an accurate round from the force exerted on the target face by a round impacting it. However, as hundreds of rounds impact in similar locations, holes form that can cause the target to remain up even when it is accurately shot. Research staff ensured all targets were refaced at the start of each training day; however, in a more precise experimental environment this variable should be more tightly controlled. If we consider this error and remove data points corresponding to where percentage exposure is significantly higher than the SPEMS score (indicating possible target issues) we see an, $R^2 = 0.64 / 0.60$. Because there was no way to prove the target was malfunctioning during these times without interrupting and altering the training environment, these results were not included in the results section. Further testing should be conducted that ensures the reliability of targets to more accurately measure the relationship between SPEMS and this MOP.

Regardless of the error introduced by the targets, SPEMS scores demonstrated significantly more predictive strength then PECL scores in terms of accuracy and consistency. Run 2 PECL data could not be modeled because it did not adequately meet the linearity, equal variance, or independence assumptions for linear regression (see Appendix D). In contrast, SPEMS data was not only a stronger predictor of percent exposure, but it was also consistent across all runs for all assumptions and conditions. The

consistency of SPEMS further indicates its strength as an assessment tool of performance. The underlying binary measurement system of the PECL makes it poorly suited to modeling performance data accurately or consistently. Further testing should be conducted to ensure the generalizability of SPEMS beyond this specific task, but results demonstrate that SPEMS is more reliable, accurate, and consistent performance measurement system then PECL in this case.

B. RECOMMENDATIONS

SPEMS scores demonstrated more predictive strength then PECL scores on task performance illustrating a more viable method for evaluating Marine Corps tasks. Following generalizability testing, SPEMS has the potential to impact at least two specific areas of the Marine Corps: training and acquisitions.

1. The Training Domain

In the training domain, SPEMS has the potential to provide training developers valuable insights into how and why their training audience is succeeding or failing at performing assigned tasks. PECLs lack any overarching method for aggregating performance step "Go/No Go" evaluations into an overall task evaluation. The underlying model provides vague standards which are ultimately evaluated qualitatively. As a result, task performance cannot be analyzed for optimization purposes because overall task performance evaluations are poorly related to specific performance step evaluations. This weak relationship can be seen in the individual performance step data shown in Chapter V Section C Table 4. Results show that improvements between runs 1 and 2 in SPEMS scores can be quickly and easily linked to improvements on specific performance steps. In contrast to the inconsistent PECL data, performance data can allow the training staff to focus future deliberated practice on these highlighted areas of weaker performance. Training developers lack the quantitative data to prove there are inefficiencies in the training continuum thus making it more difficult to justify the reallocation of resources to better train marines.

Ultimately, SPEMS has the potential to provide the accurate performance data that PECLs cannot. We recommend that SPEMS be used to evaluate marines conducting training to more accurately determine why tasks are being performed poorly, target

remediation at specific performance steps rather than the task as a whole, and to optimize current training methods. Additionally, SPEMS should be paired with video capture software to tag task performance videos with SPEMS scores. By tagging videos the Marine Corps has the potential to build large repositories of video data that can be used for further analysis. These videos could provide evaluators examples of task performance at a variety of proficiency levels that would increase rater reliability even more. Furthermore, SPEMS score tags at specific time points could be used to determine the behaviors that are linked with these scores. This type of data may allow training developers insights into what behaviors correspond to optimal task performance and may even be leveraged to build virtual evaluations. The analytical improvement of the training domain is directly related to how SPEMS can be used to inform the acquisition of training systems.

2. The Acquisition Domain

a. Training Effectiveness Evaluation Gaps

In the acquisition domain, SPEMS provides the quantitative data for evaluating how a training system supports the improvement of a marine's performance. As discussed in Chapter III section F, performance data could allow training capability developers to conduct side-by-side comparisons of training solutions to determine which solution provides the optimal mix of cost avoidance and skill acquisition. Currently, developers are able to determine the life-cycle costs of a system, estimate the costs the system avoids through simulation, and decide which training and a readiness (T&R) skills can be trained using a training effectiveness evaluation (TEE). The problem with this data is that it does not take into consideration the actual performance improvement gained by the users. The reason this problem was not previously considered is because the Marine Corps does not currently have a quantitative performance evaluation system. Training systems like the ISMT were adopted that should have avoided costs while providing equivalent training, but these systems often failed to deliver (Yates, 2004). This failure lies in the difference between a system's theoretical ability to support the training of a task, and the reality of how well the system supports the training of a task.

Training effectiveness evaluations determine what tasks a system is theoretically capable of supporting but fail to determine how well the system actually supports the training of said tasks. For example, the indoor simulated marksmanship trainer (ISMT) and live-fire marksmanship training have both been deemed capable of training marksmanship, but one may foster a more productive environment for learning the skill. In a TEE, because both methods are capable of teaching marksmanship, they will be evaluated as equally effective. The previously mentioned START process attempts to mitigate this by coding how well a system can support the training of a task on a one through five scale (Dunne, 2017). However, these improved rating systems are made by SMEs evaluating the system itself and are not based on trainees' performance improvements that result from practicing in the system. Cost avoidance data is then calculated by examining each task the system can support (evidenced by the TEE) and summing the costs associated with practicing each task in a live environment (Dunne, 2014). This method implies an equivalence between a live-fire training resource (such as a bullet), and a simulated resource.

The problem with this method is that there is no performance data to justify the assumption that the costs avoided by practicing in the proposed systems are directly proportional to the costs incurred by deliberately practicing the task live. For example, the training value of shooting a live bullet through a weapon is not the same as the training value of clicking a mousepad to shoot a weapon in a game. To determine the relationship between these alternatives, an experiment that compares the performance of participants before and after using current and proposed training support methods is required. We recommend that SPEMS be used as a method for testing the integration of a proposed training system by evaluating marine performance before and after training in each alternative. This evaluation could produce a performance discount factor (Pf) according to the following equation that would ensure cost avoidance data was proportional to the performance benefit of the system.

$$Performance\ Discount\ Factor\ (Pf) = \frac{Proposed_{Final\ SPEMS} - Proposed_{Init\ SPEMS}}{Current_{Final\ SPEMS} - Current_{Init\ SPEMS}}$$

Next, the avoided costs could be modified by multiplying the total avoided costs by the performance discount factor. This modification would ensure that the avoided costs

associated with training in the proposed environment are proportional to the comparative performance benefits it supports. This relationship is shown in the following equation.

$$Adjusted\ Avoided\ Costs = (Avoided\ Costs) * Pf$$

These properly adjusted cost avoidance estimates could then be compared to the upfront and recurring costs of fielding the proposed training solution to more accurately justify pursuing its deployment. We will walk through an example to demonstrate the importance of this proportionality.

b. Performance Discount Factor Implications Example

We will call the current training system, Live Fire Training System (LF), and the proposed training system, Integrated Simulation Training System (ISTS). Suppose that ISTS has an annual cost of $10M that is made up of the amortized upfront cost and any recurring costs. The current annual cost of conducting training for the Marine Corps in LF is $25M that is made up of practice ($15M) and qualification ($10M). The ISTS is designed to replace the practice portion of LF. Therefore, assuming the TEE determines that ISTS is capable of supporting the same training, it has the potential to avoid $15M in training costs. The annual cost of LF and ISTS is $25M + $10M = $35M with a potential cost avoidance of $15M. As a result, the minimum annual cost of the integrated system is $35M - $15M = $20M with a potential cost savings of $5M a year. This determination may lead to the integration of ISTS. However, if we were to conduct a side-by-side comparison we may see a different result.

Supposed that a control training audience will only practice in LF, and the experiment audience will practice in ISTS. Both groups are tested in the given training task prior to practicing, and their mean SPEMS score is a 1.0 at the start. After the experiment group practices in ISTS, they receive a mean SPEMS score of 2.0. After the control group practices in LF they receive a mean SPEMS score of 3.0. Therefore, the performance discount factor of the proposed system's cost avoidance would be

$$Performance\ Discount\ Factor\ (Pf) = \frac{Proposed_{Final\ SPEMS} - Proposed_{Init\ SPEMS}}{Current_{Final\ SPEMS} - Current_{Init\ SPEMS}} = \frac{2.0 - 1.0}{3.0 - 1.0} = \frac{1}{2}$$

Next, we multiply this Pf by the avoided costs of $15M for an adjusted avoided cost of $7.5M.

$$Adjusted\ Avoided\ Costs = (Avoided\ Costs) * Pf = \$15M * \frac{1}{2} = \$7.5M$$

If the total annual cost to train in ISTS and LF is $35M, with the potential to avoid $7.5M, we see a different story. The minimum future costs associated with training in the ISTS and LF are $35M - $7.5M = $27.5M with a future *added cost of at least* $2.5M. In essence, because ISTS could only accomplish 50% of the performance improvement that would have been associated with live-fire practice, it only has the potential to avoid 50% of the costs.

c. Recommendations

The conduct of the above analysis would ensure that all training systems that are fielded enable the most training value. More specifically, they support the best mix of increased performance while also avoiding the most costs. It should be noted that this type of analysis is currently done, but without any notion of a performance discount factor because performance benefits are not measured accurately (Dunne, Cooley, & Gordon, 2014). *Performance data is the missing link that aligns a proposed training system's cost avoidance proportionally with the current training method's costs. Mandating this proportionality on a performance basis is essential to actually realizing the avoided costs through usage*. Otherwise, proposed systems will be underutilized and the training (and associated costs) that were supposed to be avoided by fielding them, will not be. Users will flock to the training solution that supports the largest performance improvement to which they have access. If the alternate solution cannot compete on this metric, then it will not compete on cost avoidance. SPEMS has the potential to ensure that live, virtual, and constructive simulation is integrated to support the maximization of performance benefits while avoiding the most training costs.

C. FUTURE WORK

The success of SPEMS in this test case provides the groundwork for further investigation into SPEMS' ability to evaluate performance across all tasks and missions.

This thesis should serve as a verifiable, repeatable, and reliable proof of concept study that can be used to guide the generalizability of SPEMS. SPEMS future development should take place over four future milestones.

1. Generalize SPEMS Usability across Tasks and Missions by Conducting Multiple Proof-of-Concept Experiments to Demonstrate its Generalizability

SPEMS was developed as a generalized set of anchors that layer on top of existing training and readiness (T&R) tasks to leverage the Marine Corps' thorough decomposition of tasks by military occupational specialty (MOS) and mission. This thesis provides a proof of concept that SPEMS provides more predictive strength for evaluating buddy rush performance but is not necessarily applicable beyond the current task. To assess the generalizability of SPEMS, an experiment that determines if SPEMS' predictive strength in assessing task performance is consistent across multiple types of tasks should be conducted.

If the predictive strength of SPEMS is consistent, it would be a generalized performance evaluation measurement system that can be used across a wide range tasks and missions. If SPEMS' predictive strength was inconsistent, we recommend the Marine Corps develop a set of behaviorally anchored rating scales (BARS) specific to every MOS in the Marine Corps and re-test each new MOS specific SPEMS for generalizability within that MOS. If SPEMS' predictive strength is still inconsistent, then we recommend the Marine Corps develop a set of BARS specific to every task. This process could be done by repeating the methodologies laid out in this thesis by each community for every task. While time consuming, we feel this process would provide a quantitative, reliable, and accurate performance measurement system for each task.

2. Empirically Test a Proposed Training System Integration in a Side-by-Side Experiment with Current Training Programs to Demonstrate Performance and Cost Avoidance Advantages

Once SPEMS has been generalized across all tasks and missions, a proposed training system should be evaluated in a side-by-side comparison with current training methods to determine how each support performance benefits. This comparison would

serve as a proof of concept study to demonstrate how a proposed training system's ability to more effectively support performance improvements is evaluated. To determine whether the integration of a proposed training system provides a tangible performance benefit, the following hypothesis should be tested:

- H_0: There is no difference in the performance benefits that result from deliberate practice in the proposed (experiment) and current (control) training methods.
- H_A: There is a difference in the performance benefits that result from deliberate practice in the proposed (experiment) and current (control) training methods.

If the null hypothesis were retained, avoided costs could be assumed to be directly proportional to the costs of current training practices. Training solutions in this case would compete solely on a cost avoidance basis. If the null hypothesis were rejected, a performance discount factor would need to be applied to cost avoidance data to ensure its proportionality to performance. Either way, future researchers could move on to milestone 3 to determine if LVC simulation solutions should be integrated into current training methodologies.

3. Combine Performance Data with Life-Cycle Cost and Cost Avoidance Data to Determine the Return on Investing in Proposed Training Programs

This milestone was largely discussed in the Recommendations section, sub-section 2, but would entail the linking of performance data and cost data through a performance discount factor (Pf) to ensure a proportional relationship. Currently, proposed training solutions' avoided costs are calculated by summing the costs of equivalent live training. However, there are performance benefits associated with live training that are not captured in simulation due to a degradation in fidelity. By using this directly proportional relationship, we are knowingly introducing inaccuracies into this calculation. Instead, cost data must be discounted according to the Pf calculated by a side-by-side study of training

solutions. The accuracy of this relationship could only be tested in milestone 4 by comparing the estimated returns with realized returns.

4. Monitor the Integrated Training Solution throughout its Life Cycle to Determine the Accuracy of Current and Revised ROI Calculations

Once the proportion between the proposed solution's avoided costs and current training solution's costs is accurate, estimated avoided costs would have to be compared to reality to ensure the model is accurate. If the performance discount factor's incorporation into cost avoidance modeling proved to be more accurate over time, its value would be validated. This type of return on investment calculation could be accepted as a validated method for evaluating proposed training system integration. If the performance data's incorporation proved to be less accurate over time, the relationship between performance and cost data would have to be reviewed and redefined. This final milestone would have to be retested until a validated model was produced that accurately linked cost avoidance and performance data. The successful retention of the above null hypothesis would indicate that the training systems getting fielded provide a return on investment from both a cost and performance perspective. This would indicate that a system's overall benefits were properly being captured and used to inform the integration of live, virtual, and constructive technologies into the training and readiness program.

D. SUMMARY

This thesis focused on developing a scaled performance evaluation measurement system (SPEMS) that evaluates how Marines perform tasks to quantitatively demonstrate the benefits of integrating new training programs. The primary problem with evaluating a training program's performance benefits is the current binary performance evaluation system that does not provide the structure to quantitatively measure performance. We addressed this problem by examining how task proficiency is developed through deliberate practice, how the measurement and analyses of these processes could be improved, and what quantitative task evaluation techniques currently exist. Based on this background research, we developed SPEMS through an iterative process that began with a card-sorting task. The initial set of BARS were refined iteratively by convening three SME focus groups

to measure the reliability of SPEMS. SPEMS proved to be an extremely reliable performance evaluation tool (Cronbach's Alpha 0.93 - 0.98); therefore, we conducted an empirical evaluation of SPEMS' accuracy in the operational environment. This empirical evidence allowed us to reject our null hypothesis and conclude that SPEMS (R^2 = 0.41/ 0.40) has more predictive power than current PECL (R^2 = 0.20) techniques on objective measures of buddy rush task performance. Finally, we recommend that SPEMS 1) be empirically generalized across tasks and missions; 2) be used to empirically test a proposed training system integration in a side-by-side experiment with current training programs to demonstrate performance and cost avoidance advantages; 3) be combined with cost data to more accurately determine the return on investing in LVC integration; and 4) Be validated through the monitoring of actual cost data compared to the estimated cost models. These findings demonstrate a verifiable, repeatable, and reliable potential solution to the problem of measuring military task performance across training solutions.

THIS PAGE INTENTIONALLY LEFT BLANK

APPENDIX A. THE TRAINING AND READINESS TASK

Appendix A describes the components of a training and readiness (T&R) task. The T&R task consists of a number of elements that all play a significant role in conducting and monitoring the progress of military deliberate practice. Figure 20 shows the task, "INF-MAN-3001: Conduct Fire and Movement," that was featured in this thesis.

INF-MAN-3001: Conduct fire and movement

SUPPORTED MET(S):
MCT 1.14 MCT 1.6.1 MCT 1.6.4

EVALUATION-CODED: NO **SUSTAINMENT INTERVAL**: 6 months

CONDITION: Given an order from higher and an enemy.

STANDARD: To neutralize the enemy threat in order to accomplish the mission, meeting the commander's intent.

EVENT COMPONENTS:
1. Suppress the enemy (S).
2. Assess effects of fires (A).
3. Adjust fires as necessary.
4. Identify next covered position.
5. Move to next covered position under the cover of suppression(M).
6. Identify your target and continue suppression to allow buddy to move to next covered position.
7. Repeat steps 1-5 until the objective is reached.
8. Execute actions on the objective (K).
9. Consolidate.

REFERENCES:
1. FM 21-75 Combat Skills of the Soldier
2. MCRP 3-10A.1 Infantry Company Operations
3. MCRP 3-10A.3 Marine Rifle Squad

INTERNAL SUPPORTING EVENTS:
0300-PAT-1005 0311-OFF-2001

INTERNAL SUPPORTED EVENTS: INF-MAN-4001

SUPPORT REQUIREMENTS:

SIMULATION EVALUATION:

SIMULATED	SUITABILITY	SIMULATOR	UNIT OF MEASURE	HOURS	PM
Yes	S/L	I-TESS (Individual)	Marine Hours	2	N

ORDNANCE:

DODIC		QUANTITY
A059	Ctg, 5.56mm Ball M855 (Clip)	40 rounds per Marine
A063	Ctg, 5.56mm Tracer M856	10 round per Marine
A080	Ctg, 5.56mm Blank M200	50 round per Marine
B504	Ctg, 40mm Green Star Para M661 F/M20	1 cartridges per Team
B508	Ctg, 40mm Green Smoke M715 F/M203 Gr	1 cartridges per Team
B509	Ctg, 40mm Yellow Smoke M716 F/M203 G	1 cartridges per Team
B535	Ctg, 40mm White Star Para M583 F/M20	1 cartridges per Team
B546	Ctg, 40mm HEDP M433 F/M203 Gren Laun	2 cartridges per Team
BA35	Ctg, 40mm Prac Low Velocity	5 cartridges per Team
C995	Launcher & Ctg,84mm M136 (AT-4)	1 cartridges per Team
G811	Body, Practice Hand Grenade M69	1 grenades per Team
G878	Fuze, Hand Grenade M228	1 fuze per Team
G881	Gren, Hand Fragmentation M67	1 grenades per Team
G940	Gren, Hand Smoke Green M18	1 grenades per Team
G945	Gren, Hand Smoke Yellow M18	1 grenades per Team
G955	Gren, Hand Smoke Violet M18	1 grenades per Team
G982	Gren, Hand Smoke TA M83	1 grenades per Team
HA21	Sub-cal Rkt, Practice 21mm	1 rockets per Team
HA29	Rckt 66mm HE, M72A7, LAW W/GRAZE	1 rockets per Team
L594	Sim, Proj Ground Burst M115A2	1 Simulator per Team

RANGE/TRAINING AREA:
Facility Code 17410 Maneuver/Training Area, Light Forces
Facility Code 17430 Impact Area Dudded
Facility Code 17730 Fire And Movement Range
Facility Code 17750 Infantry Squad Battle Course

MISCELLANEOUS:

ADMINISTRATIVE INSTRUCTIONS: Considerations, means of movement include unit, buddy, and individual. The event may also be used for cover and movement when there is no immediate enemy threat. A leader issues the ADDRAC in support of this event.

Figure 20. The task "INF-MAN-3001: Conduct Fire and Movement." Source: Marine Corps (2016a).

Event Code: All events in T&R manuals are either individual or collective and are made up of four sets of four numbers. The first set indicates the MOS associated with the event (0311 for infantryman). The second set indicates the functional or duty area (PAT for patrolling). The third set indicates the level and sequence of the task. 1000–2000 level tasks are individual tasks, and 3000–8000 level tasks are collective tasks that range from the fire team to the regimental level. The higher the task number is within a current level, the more advanced the task is (3001 should be completed prior to 3101) (Marine Corps, 2011).

Title: This is the name of the event or task.

Evaluation Code: Evaluation coded events are directly linked to mission essential tasks (MET). These events are formally evaluated and must be included in a units training program (Marine Corps, 2011). They are known as E-Coded events.

Supported MET(s): If a task is 'E-Coded' this is where all associated METs are listed.

Sustainment Interval: This is the period expressed in number of months between evaluation (Marine Corps, 2011).

Billet/MOS: This is the recommended MOS responsible for conducting the task. It is important to note that the commander has the operational flexibility to shift this responsibility if he feels another MOS is more capable or poised to accomplish the task.

Grade: The rank of Marines required in order to complete the event.

Description: The description provides a short explanation of the purpose of the event.

Condition: Conditions are the constraints each Marine must abide by for the event to be trained to a given standard. It indicates the physical conditions required and the equipment the Marines are authorized to use. For example, wearing a fighting load, during day/night, during cold weather conditions…etc. Commanders are able to modify these conditions for safety or operational factors that may limit a units' ability to conduct the task as prescribed.

Standard: The standard is the focus of this thesis and, "indicates the basis for judging the effectiveness of the performance. It consists of a carefully worded statement that identifies the proficiency level expected when the task is performed" (Marine Corps, 2011, p. 4–3). The standard provides the lowest level of performance required to be considered trained in the given task and can range from specific quantitative metrics for individual events to general statements for collective events. The example standard in Figure 1, "to accomplish the mission and meet commander's intent," provides no quantifiable metrics or measures of performance and calls into question the Marine Corps' statement that it is a, "carefully worded statement that identifies the proficiency level expected" (Marine Corps, 2011, p. 4–3).

Event Components/Performance Steps: The performance steps spell out the subordinate tasks or actions required to complete the event to standard. Event components are named for collective events with subordinate tasks, and performance steps are used for individual events (Marine Corps, 2011).

Prerequisite Events: These are the T&R events that must be completed prior to attempting the given task.

Chained Events: Higher level tasks are chained to lower level tasks which enables commanders to identify subordinate events that support the performance of the mission essential task list. Each task lists any lower level tasks that are chained to the upper level task (Marine Corps, 2011).

Related ITSs: A list of all tasks in the sequence prior to the event that are related or support the given event.

Initial Training Setting: All individual events list where the task is initially trained. This is typically done at formal school (FS) or in the operational forces.

References: The references section links every ITS to the doctrinal, warfighting, or reference publication that substantiates the task. These publications provide further guidance, historical examples, and support for how to conduct and train the tasks. The T&R provides training guidelines that are meant to be further substantiated by these publications to build well rounded training programs (Marine Corps, 2011).

Distance Learning/Simulation Products: This indicates if the event is capable of being trained using some modality other than live training and what the modality is.

Support Requirements: This is a list of all of the internal/external support required for units to train the task. These requirements typically include weapon systems, equipment, training ranges, and other materials (Marine Corps, 2011).

Miscellaneous: Any additional information that supports the training of the task.

THIS PAGE INTENTIONALLY LEFT BLANK

APPENDIX B. SPEMS SURVEY

Appendix B includes a copy of the SPEMS survey that was given to SME participants of the virtual video analysis focus groups.

Scaled Performance Evaluation Measurement System (SPEMS) Effectiveness Survey

LEAST **MOST**

How easy was SPEMS to use?

1 2 3 4 5 6 7 8 9 10

Do you think SPEMS is an effective tool for measuring training?

1 2 3 4 5 6 7 8 9 10

How reliably does SPEMS measure task performance between evaluators?

1 2 3 4 5 6 7 8 9 10

Do you think SPEMS is _____ than traditional PECL evaluation methods?

Less Effective **As Effective** **More Effective**

How much training would it require for evaluators to use SPEMS?

0 10 20 30 40 50 60 **Minutes** > 60:_____

How much training is required for SPEMS to be used reliably between evaluators?

0 10 20 30 40 50 60 **Minutes** > 60:_____

Please rate the anchors associated with each level according to their ability to represent performance at that level?

1 - Novice: Unable to execute. Performance step not addressed. No acknowledgement.

1 2 3 4 5 6 7 8 9 10

2 - Advanced beginner: Performance step attempted, majority mistakes. Below standard.

1 2 3 4 5 6 7 8 9 10

3—Competent: Performance step attempted, some mistakes.

1 2 3 4 5 6 7 8 9 10

4—Proficient: No references/guidance required. Executed to standard. Few mistakes.

1 2 3 4 5 6 7 8 9 10

5 - Mastery: Flawless Execution. Performance step completed, no mistakes.

1 2 3 4 5 6 7 8 9 10

What did you like about SPEMS?

How do you think SPEMS could be improved?

Do you think SPEMS videos can be used for baselining evaluators on task performance levels?

Questions, Comments or Concerns about SPEMS?

APPENDIX C. MEASURES OF PERFORMANCE SURVEY

Appendix C includes a copy of the final objective measure of buddy rush performance survey that was given to SME participants in the virtual video analysis focus groups.

Conduct Fire and Movement: Objective Measures of Performance Survey

What objective measures of performance do you think most accurately measure performance in this task?

_____

How long have you served? _____ years

What is your rank? _____

Do you think you are qualified to evaluate performance of this task? **Yes** **No**

How familiar are you with the buddy rush task?

LEAST **MOST**

1 2 3 4 5 6 7 8 9 10

Are any of the following a measure of performance for this task?

Accuracy of rounds fired? **Yes** **No**

Number of individual rushes? **Yes** **No**

Number of target bobs per minute? **Yes** **No**

Rate of fire? **Yes** **No**

Time spent moving with target up? **Yes** **No**

Number of times cover not used? **Yes** **No**

Rank order from Most important (1) to Least Important (5)

_____ Accuracy of rounds fired?

_____ Number of individual rushes?

_____ Number of target bobs per minute?

_____ Rate of fire?

_____ Time spent moving with target up?

_____ Number of times cover not used?

THIS PAGE INTENTIONALLY LEFT BLANK

APPENDIX D. ASSUMPTIONS AND CONDITIONS

Appendix D details how the assumptions and conditions for every statistical test used within this thesis were addressed, checked, and met. They are laid out in the same order that the statistical tests are presented in the thesis.

A. PART 1: SCORING—PAIRED T-TEST FOR 2 SAMPLE MEANS

The following runs were compared and tested using a paired t-test for 2 sample means. The assumptions and conditions for a paired t-test for 2 sample means are: sample size, paired data, continuous data, and normality. There is an n = 26 buddy pairs which is larger than the minimum for a paired t-test, n = 15. The same 26 buddy pairs were evaluated three times, so comparing runs is paired by buddy pair in each test. SPEMS scores are continuous between the interval 0–5 and PECL scores are continuous on the interval between 0–1. Percent Exposure is continuous between the interval 0 to 100. Normality for each test was demonstrated with a histogram of the differences between runs for each buddy pair.

1. SPEMS Score Means

a. Run 1 & 2: Paired t-test for 2 Sample Means

The histogram in Figure 21 is left skewed and approached non-normality evidenced by the Shapiro-Wilk goodness of fit test (p = 0.0523). As a result, a Wilcoxon signed rank test was used in the thesis to demonstrate a difference between run 1 and run 2 for SPEMS scores.

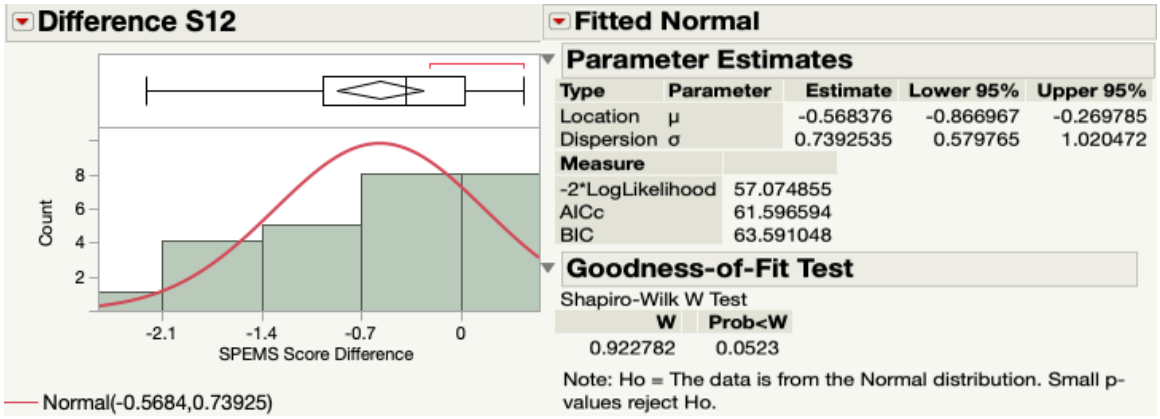


Figure 21. Histogram of the differences in SPEMS scores between run 1 and 2 with a fitted normal curve

b. Run 2 & 3: Paired t-test for 2 sample means

The histogram in Figure 22 is approximately normal with a slight right skewness. The distribution meets the normal enough assumption for a paired t-test.

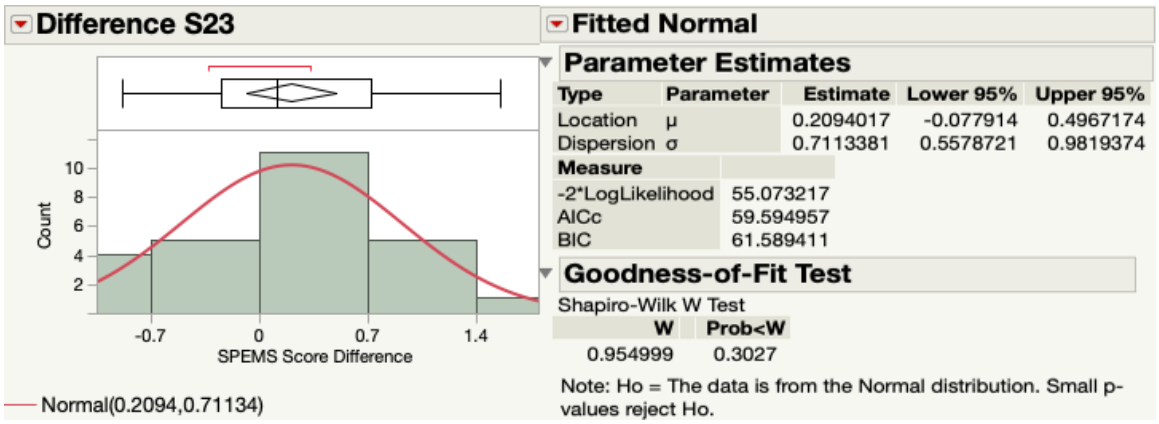


Figure 22. Histogram of the difference in SPEMS scores between runs 2 and 3 with a fitted normal curve

2. PECL Score Means

a. Run 1 & 2: Paired t-test

The histogram in Figure 23 does not meet the normality assumption for a paired-t test. The Shapiro-Wilk test for goodness of fit shows that the distribution is not normal

enough and as a result, a Wilcoxon signed rank test was used in the thesis to demonstrate a difference between run 1 and run 2 for PECL scores.

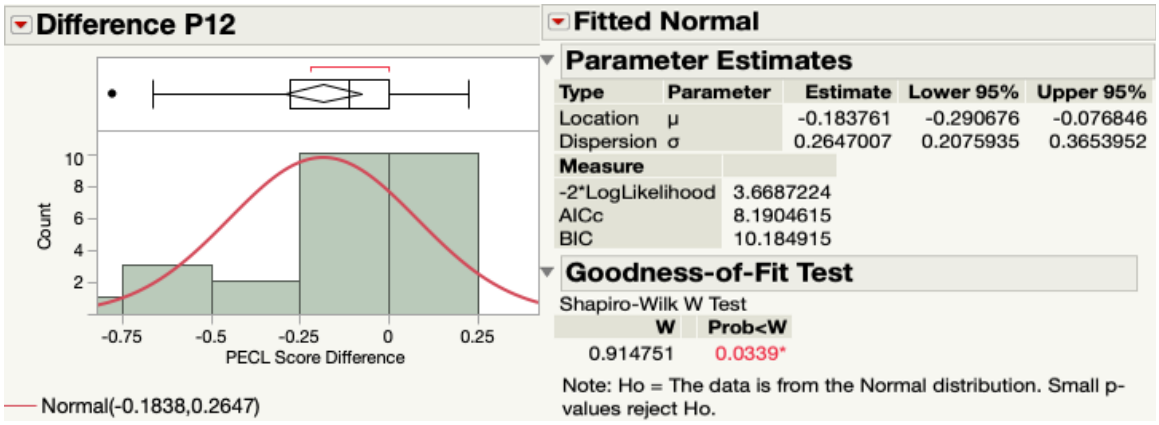


Figure 23. Histogram of the difference in PECL scores between run 1 and 2 with a fitted normal curve

b. Run 2 & 3: Paired t-test

The histogram in Figure 24 is approximately normal. The distribution meets the normal enough assumption for a paired t-test.

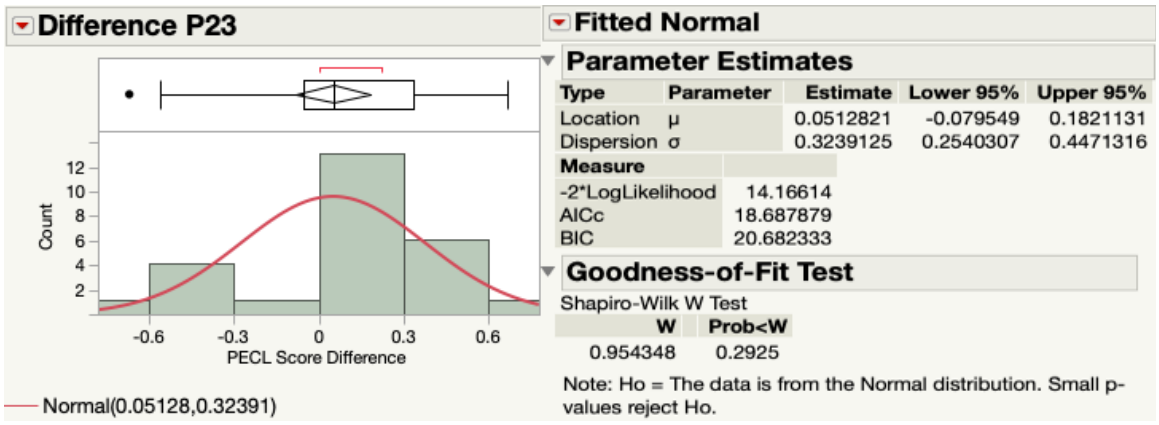


Figure 24. Histogram of the difference in PECL scores between run 2 and 3 with a fitted normal curve

3. Mean Percent Exposure

a. Run 2 & 3: Paired t-test for 2 Sample Means

The histogram in Figure 25 is approximately normal. The distribution meets the normal enough assumption for a paired t-test for 2 sample means.

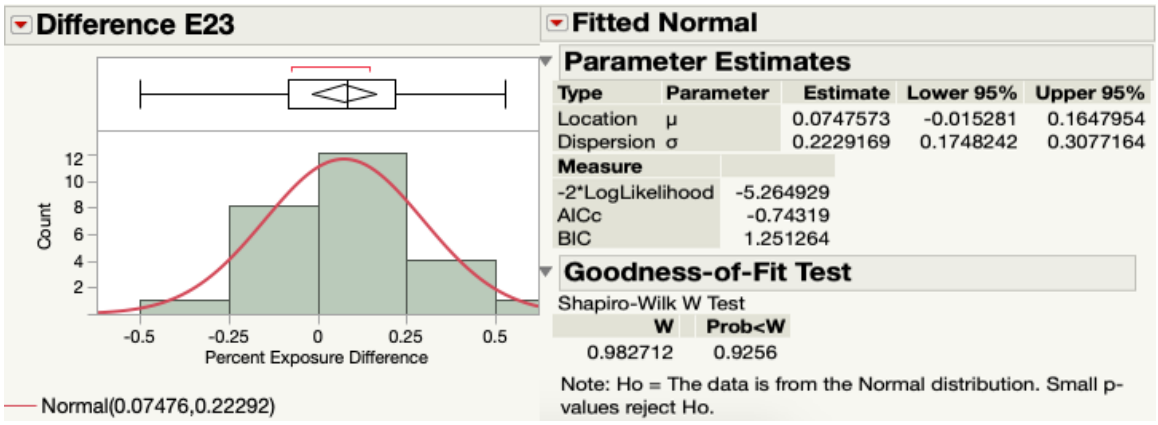


Figure 25. Histogram of the difference in percent exposure between runs 2 and 3 with a fitted normal curve

B. PART 2: PERCENT EXPOSURE AND SCORING—LINEAR REGRESSION

SPEMS and PECL scores were tested for having a relationship with percent exposure using linear regression. The assumptions and conditions for linear regression are independence, linearity, normality, and unequal variance. These assumptions and conditions were checked the following four times for each run and each type of scoring technique. The SPEMS score data set met all assumptions and conditions for both runs. The PECL data set does not meet all assumptions and conditions.

1. SPEMS and Percent Exposure

a. Run 2

(1) Linearity

Figure 26 is a regression plot of SPEMS score by percent exposure for run 2. The plot demonstrates a strong degree of linearity as evidenced by the linearly arrayed data in a relatively strong negative relationship. The regression plot meets the linearity condition.

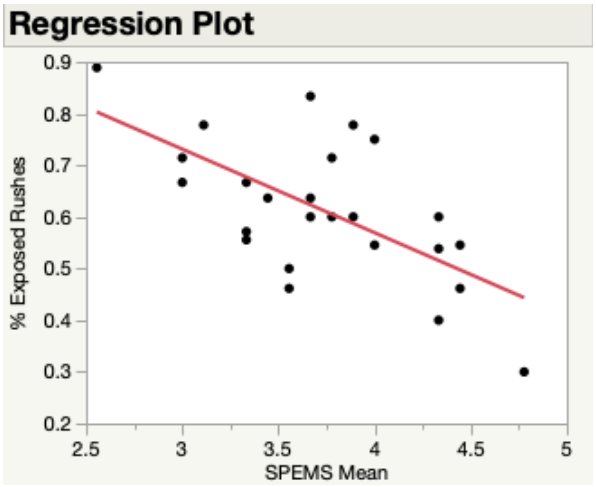


Figure 26. Run 2 SPEMS score vs. percent exposed rushes plot demonstrating linearity

(2) Normality

The histogram in Figure 27 demonstrates a normal enough distribution of the residuals that result from the linear regression. The distribution meets the normality condition.

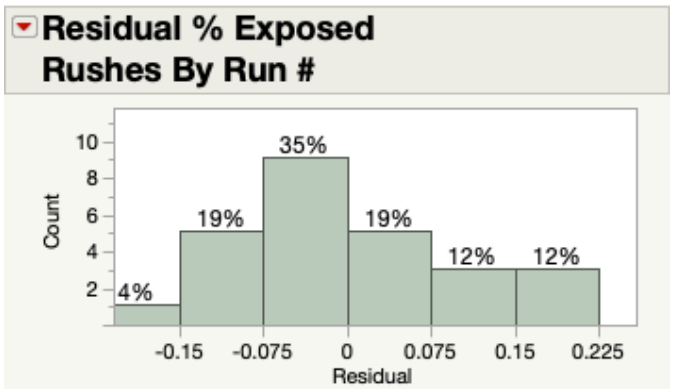


Figure 27. Histogram of residuals resulting from a linear regression between percent exposure and SPEMS score for run 2

(3) Equal Variance and Independence

The residual by predicted plot in Figure 28 demonstrates a random distribution of data points with minimal clumping or clustering. This demonstrates that the data set for SPEMS score and percent exposure for run 2 meets the equal variance and independence conditions.

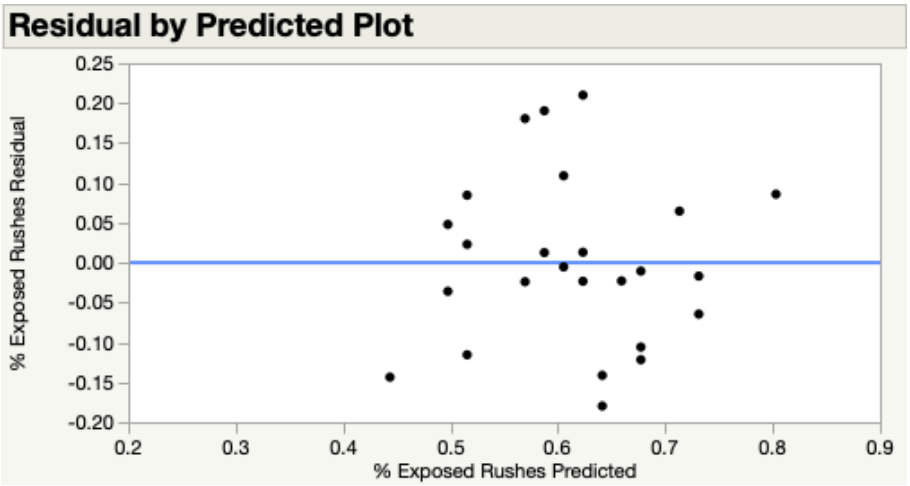


Figure 28. SPEMS residual plot for percent exposed predicted by percent exposed residuals for run 2

b. Run 3

(1) Linearity

Figure 29 is a regression plot of SPEMS scores by percent exposure for run 3. The plot demonstrates a strong degree of linearity as evidenced by the linearly arrayed data in a relatively strong negative relationship. The regression plot meets the linearity condition.

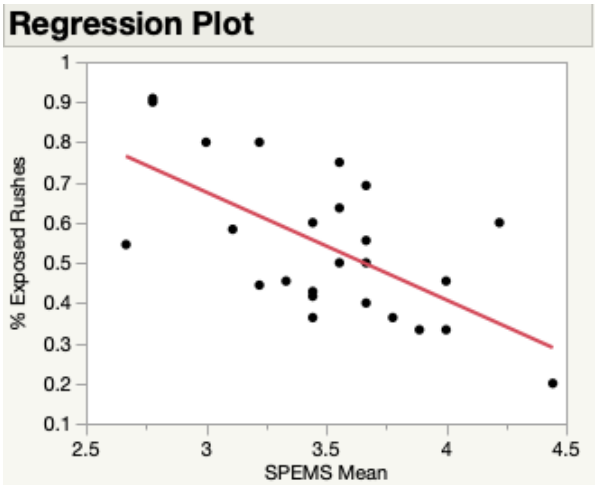


Figure 29. Run 3 SPEMS score vs. percent exposed rushes plot demonstrating linearity

(2) Normality

The histogram in Figure 30 demonstrates a normal enough distribution of the residuals that result from the linear regression. The histogram is slightly right skewed but is normal enough for proceeding. The distribution meets the normality condition.

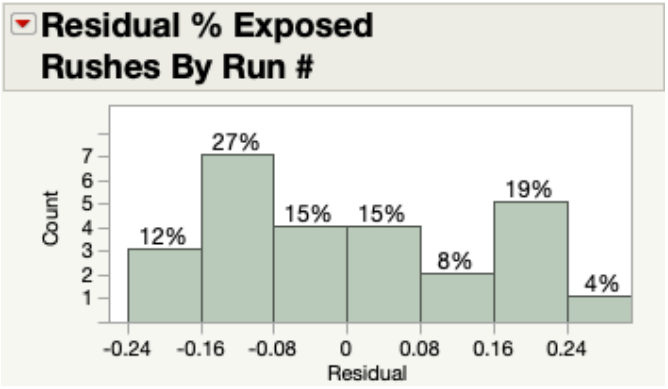


Figure 30. Histogram of residuals resulting from a linear regression between percent exposure and SPEMS scores for run 3

(3) Equal Variance and Independence

The residual by predicted plot in Figure 31 demonstrates a random distribution of data points with minimal clumping or clustering. This demonstrates that the data set for SPEMS scores and percent exposure for run 3 meets the equal variance and independence conditions.

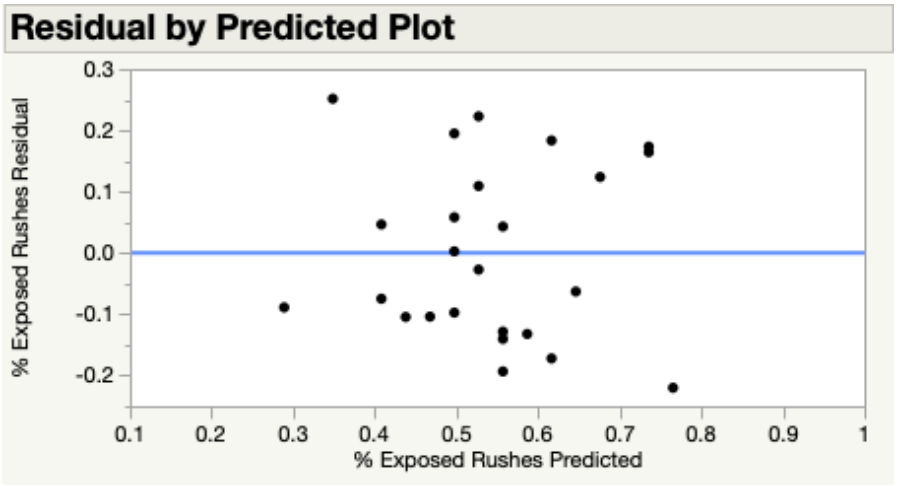


Figure 31. SPEMS scores residual plot for percent exposed predicted by percent exposed residuals for run 3

2. PECL and Percent Exposure

a. Run 2

(1) Linearity

Figure 32 is a regression plot of PECL score by percent exposure for run 2. The plot demonstrates almost no linearity as evidenced by the triangular looking distribution with clumping at 1.0 and 0.9. The regression plot does not meet the linearity condition.

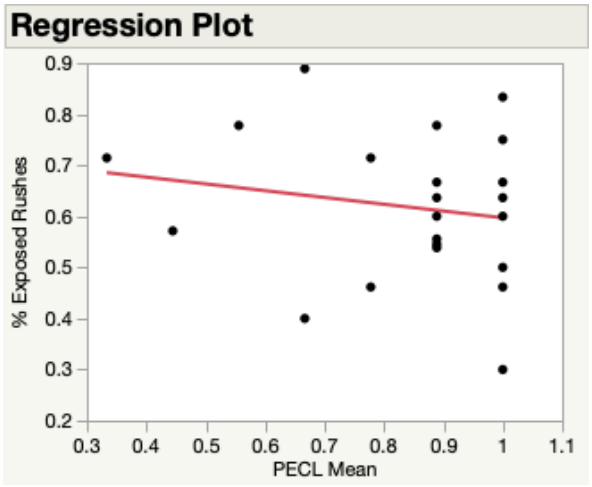


Figure 32. Run 2 PECL scores vs. percent exposed rushes plot demonstrating non-linearity

(2) Normality

The histogram in Figure 33 demonstrates a normal enough distribution of the residuals that result from the linear regression. The distribution meets the normality condition.

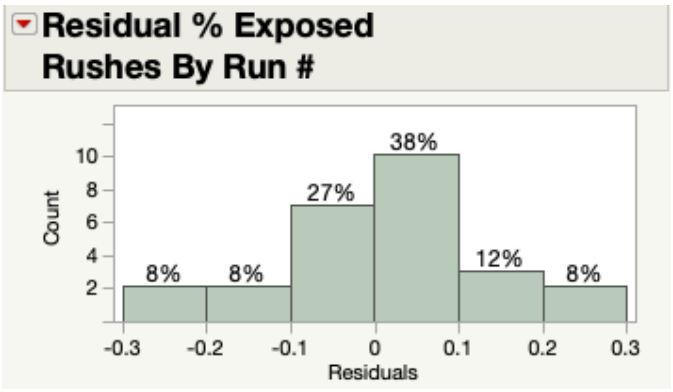


Figure 33. Histogram of residuals resulting from a linear regression between percent exposure and PECL score for run 2

(3) Equal Variance and Independence

The residual by predicted plot in Figure 34 demonstrates while there is no clumping or clustering, lack of range on the x-axis is of deep concern.

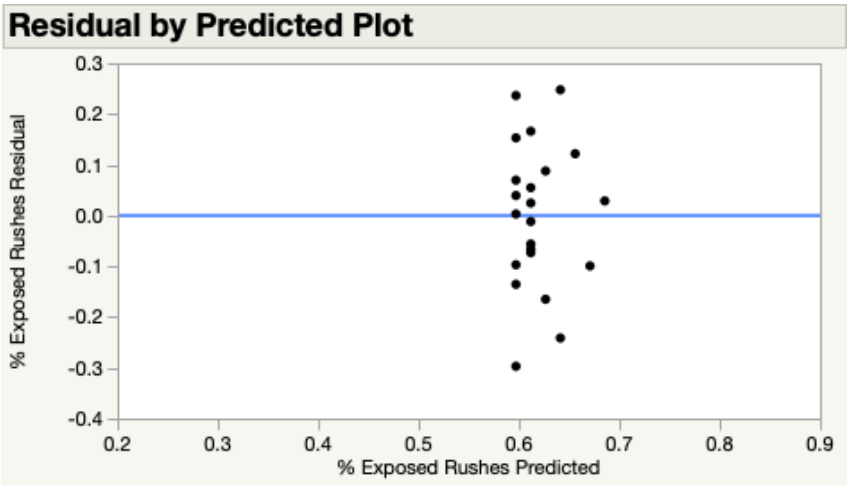


Figure 34. PECL residual plot for percent exposed rushes predicted by percent exposed rushes residuals for run 2

The concerns stated regarding linearity, equal variance, and independence all led to the assumptions and conditions not being met. The linear regression is shown in Figure 35.

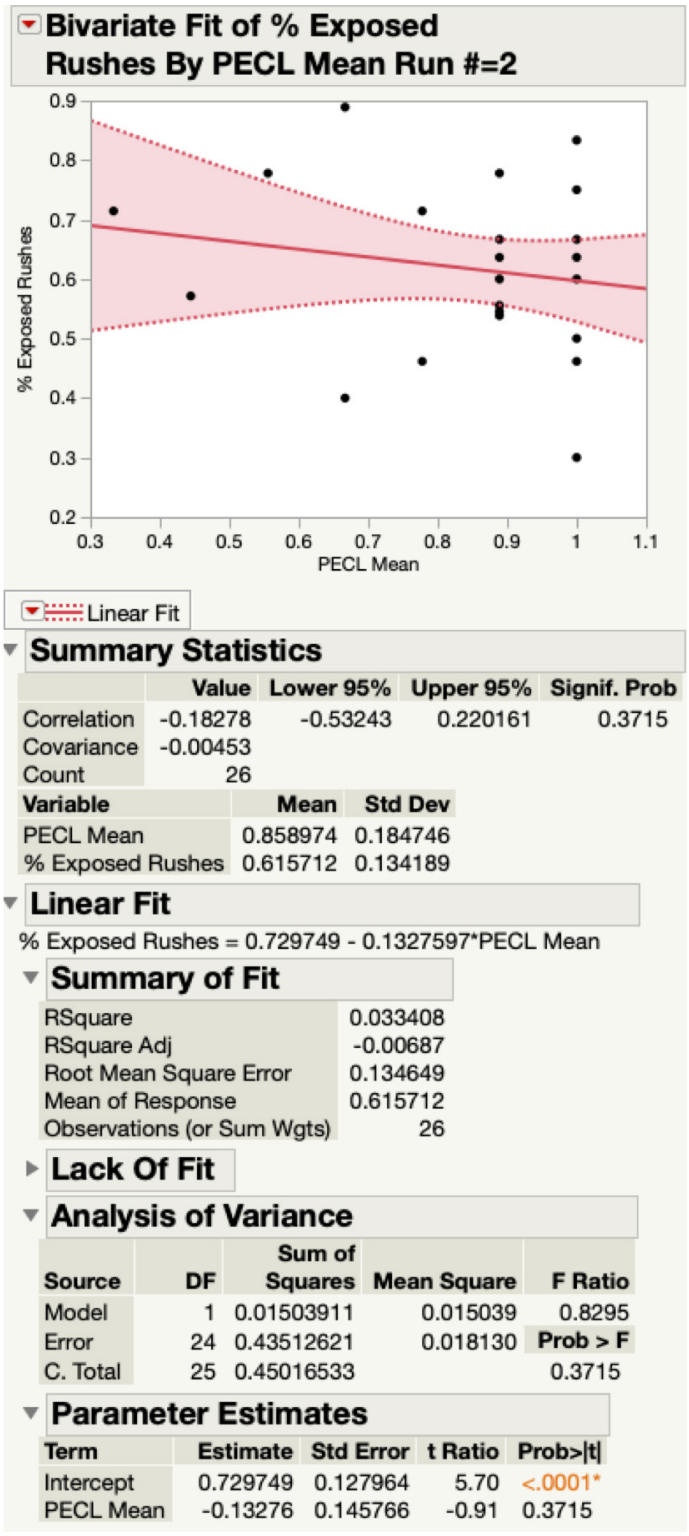


Figure 35. Linear regression results testing fit of mean PECL score to percent exposure. Run 2 was rejected and shows R^2=0.03 demonstrating no fit

b. Run 3

(1) Linearity

Figure 36 is a regression plot of PECL score by percent exposure for run 2. The plot demonstrates almost no linearity as evidenced by the triangular looking distribution with clumping at 1.0 and 0.9. The regression plot does not meet the linearity condition.

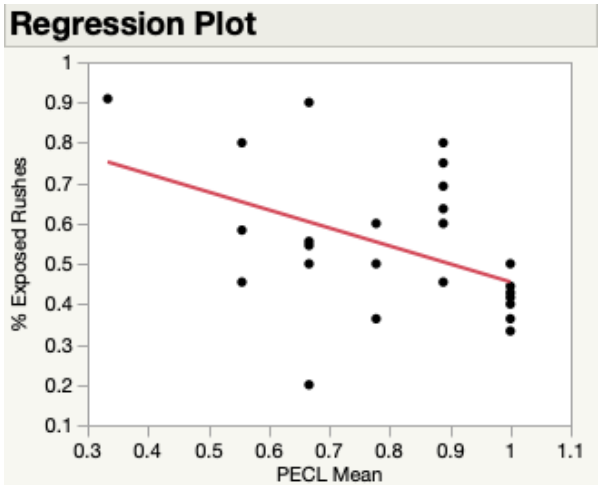


Figure 36. Run 3 PECL scores vs. percent exposed rushes plot demonstrating non-linearity

(2) Normality

The histogram in Figure 37 demonstrates a normal enough distribution of the residuals that result from the linear regression. The distribution meets the normality condition.

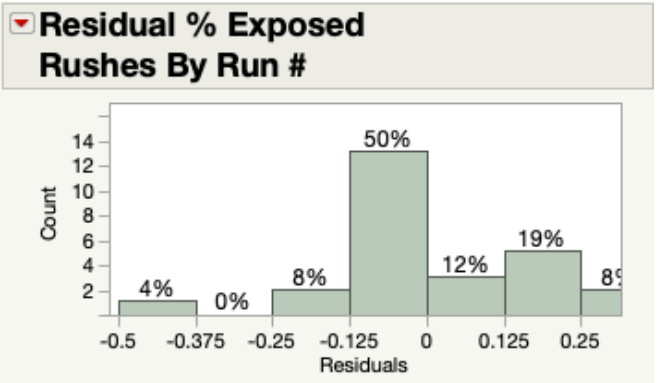


Figure 37. Histogram of residuals resulting from a linear regression between percent exposed rushes and PECL scores for run 3

(3) Equal Variance and Independence

The residual by predicted plot in Figure 38 demonstrates a significant amount of clumping and clustering. Therefore, there is concern the data set for PECL scores and percent exposure for run 3 may not meet the unequal variance or the independence condition.

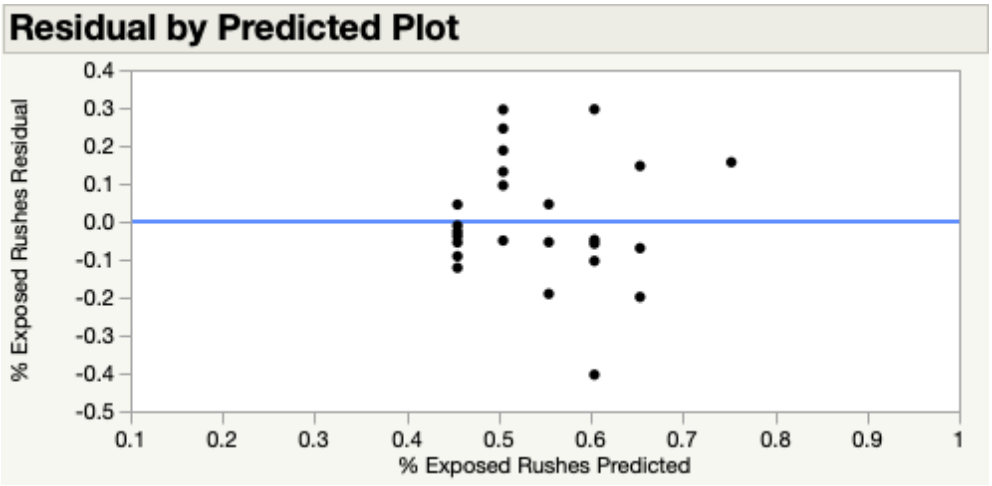


Figure 38. PECL residual plot for percent exposed rushes predicted by percent exposed rushes residuals for run 3

THIS PAGE INTENTIONALLY LEFT BLANK

LIST OF REFERENCES

Boldovici, J. A., Bessemer, D. W., & Bolton, A. E. (2002). *The elements of training evaluation*. Retrieved from https://www.researchgate.net/publication/235157151_The_Elements_of_Training_Evaluation

Brewer, N. T., & Chapman, G. B. (2002). The fragile basic anchoring effect. *Journal of Behavioral Decision Making*, *15*(1), 65–77. https://doi.org/10.1002/bdm.403

Dunne, R., Cooley, T., & Gordon, S. (2014). Proficiency evaluation and cost-avoidance proof of concept M1A1 study results (Paper No. 14055). In *Interservice/Industry Training, Simulation, and Education Conference (I/ITSEC)* (pp. 1–12). Retrieved from http://www.dtic.mil/dtic/tr/fulltext/u2/a620169.pdf

Dunne, R., Harris, S., Arrieta, A., Tanner, S., Vonsik, B., Lalor, J., & Muir, S. (2017). Live, virtual, constructive distributed missions: Results and lessons learned (Paper No. 17229). In *Interservice/Industry Training, Simulation, and Education Conference (I/ITSEC)* (pp. 1–12). Retrieved from http://www.iitsecdocs.com/search/

Evaluation. (n.d.). In *Merriam-Webster.* Retrieved August 2017 from https://www.merriam-webster.com/dictionary/evaluation

Fishburne, R., Murray, M., & Blair, A. (1979). *E-2 systems approach to training: development, implementation, evaluation and revision* (Report No. NAVTRAEQUIPCEN 78-C-0045-1). Retrieved from https://apps.dtic.mil/dtic/tr/fulltext/u2/a080428.pdf

Fletcher, J. D., & Chatelier, P. R. (2000). *An overview of military training* (Report No. D-2514). Retrieved from http://www.dtic.mil/dtic/tr/fulltext/u2/a408439.pdf

Gliner, J. A., Morgan, G. A., & Leech, N. L. (2011). *Research methods in applied settings* (2nd ed.). New York, NY: Routledge.

Government Accountability Office. (2013). *Army and Marine Corps training: Better performance and cost data needed to more fully assess simulation-based efforts*. Washington, DC: Author. Retrieved from https://www.gao.gov/assets/660/657115.pdf

Hodges, G., Darken, R., & McCauley, M. (2014). *An analytical method for assessing the effectiveness of human in the loop simulation environments: A work in progress*. (Doctoral dissertation). Retrieved from https://calhoun.nps.edu/handle/10945/46350

Hubbard, D. W. (2014). *How to measure anything: finding the value of intangibles in business* (3rd ed.). Hoboken, NJ: John Wiley and Sons.

Jones, N., Seavers, G., Capriglione, C., & Jones, N. (2015). Measuring virtual simulation's value in training exercises—USMC use case (Paper No. 15114). In *Interservice/Industry Training, Simulation, and Education Conference (I/ITSEC)* (pp. 1–12). Retrieved from https://apps.dtic.mil/docs/citations/AD1001873

Kaplan, R. M. (1978). Is beauty talent? Sex interaction in the attractiveness halo effect. *Sex Roles*, *4*(2), 195–204. https://doi.org/10.1007/BF00287500

Kingstrom, P., & Bass, A. (1981). A critical analysis of studies comparing behaviorally anchored rating scales (BARS) and other rating formats. *Personnel Psychology*, *34*, 27. Retrieved from https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1744-6570.1981.tb00942.x

Kirkpatrick, D. L., & Kirkpatrick, J. D. (2006). *Evaluating training programs: The four levels*. San Francisco, CA: Berrett-Koehler.

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

Marine Corps. (2004). *Systems approach to training*. Quantico, VA: Author. Retrieved from https://www.trngcmd.marines.mil/Portals/207/Docs/FLW/EEIC/SAT_Manual.pdf

Marine Corps. (2008) *Professional military education* (PME) (MCO 1553.4B). Washington, DC: Author. Retrieved from https://www.marines.mil/Portals/59/MCO 1553.4B.pdf

Marine Corps. (2011). *Marine Corps ground training and readiness program (*MCO P3500.72A). Washington, DC: Author. Retrieved from https://www.marines.mil/Portals/59/Publications/MCO P3500.72A.pdf?ver=2012-10-11-163735-363

Marine Corps. (2015a). *How to conduct training* (MCRP 3–0B). Quantico, VA: Author. Retrieved from https://www.marines.mil/Portals/59/Publications/MCRP 3–0B.pdf

Marine Corps. (2015b). *Performance evaluation system* (MCO 1610.7). Washingto, DC: Author. Retrieved from https://www.marines.mil/Portals/59/Publications/MCO 1610.7.pdf

Marine Corps. (2016a). *Infantry training and readiness manual* (NAVMC 3500.44C). Washington, DC: Author. Retrieved from https://www.marines.mil/Portals/59/Publications/NAVMC 3500.44C Infantry T-R Manual (secured).pdf?ver=2017-03-09-080222-740

Marine Corps. (2016c). *Unit training management guide* (MCTP 8–10A). Washington, DC: Author. Retrieved from https://www.marines.mil/Portals/59/Publications/MCTP 8–10A.pdf?ver=2017-03-16-121330-570

Morrison, J., & Meliza, L. (1999). *Foundations of the after action review process* (Report No. Special Report #42). Retrieved from http://www.dtic.mil/dtic/tr/fulltext/u2/a368651.pdf

Proficiency. (n.d.). In *Merriam-Webster*. Retrieved August 2017 from https://www.merriam-webster.com/dictionary/proficiency

Richardson, J. J. (2013). *Developing behavioral metrics for decision-making in Marine Corps small-units*. (Master's Thesis). Retrieved from https://calhoun.nps.edu/handle/10945/37701

Rubin, D. B., Cronbach, L. J., Gleser, G. C., Nanda, H., & Rajaratnam, N. (1974). The Dependability of Behavioral Measurements: Theory of Generalizability for Scores and Profiles. *Journal of the American Statistical Association 69*(348), 1050. https://doi.org/10.2307/2286194

Schwab, D. P., Heneman Herbert G., I. I. I., & DeCotiis, T. A. (1975). Behaviorally anchored rating scales: A review of the literature. *Personnel Psychology*, *28*(4), 549–562. Retrieved from http://libproxy.nps.edu/login?url=http://search.ebscohost.com/login.aspx?direct=true&db=bth&AN=6265587&site=ehost-live&scope=site

Taras, M. (2005). Assessments—summative and formative: Some theoretical reflection. *British Journal of Educational Studies*, *53*(4), 466–478. https://doi.org/10.1111/j.1467-8527.2005.00307.x

Trabun, M. A. (2007). *U.S. Marine Corps training modeling and simulation master plan*. Quantico, VA: United States Marine Corps Training and Education Command. Retrieved from https://apps.dtic.mil/dtic/tr/fulltext/u2/a471953.pdf

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science*, *185*(4157), 1124–1131. Retrieved from https://www.jstor.org/stable/1738360?seq=1/analyze

United States War Department, Fleury, F., & Steuben, F. (1807). *Regulations for the order and discipline of the troops of the United States*. Printed for Evert Duyckinck. Retrieved from http://hdl.handle.net/2027/nyp.33433008596672

Usability Professionals Association. (2010). Card Sorting | Usability Body of Knowledge. Retrieved January 3, 2019, from https://www.usabilitybok.org/card-sorting

Wigdor, A. K., & Green, B. F. (1991). *Performance assessment for the workplace* (Vol. II). Washington, DC: National Academies Press. https://doi.org/10.17226/1898

Wong, L., Gerras, S. J., & Barracks, C. (2015). *Lying to ourselves: dishonesty in the Army profession*. Carlisle, PA: Strategic Studies Insitute and U.S. Army War College Press. Retrieved from https://ssi.armywarcollege.edu/pdffiles/pub1250.pdf

Yates, W. (2004). *A training transfer study of the Indoor Simulated Marksmanship Trainer* (Master's Thesis). Retrieved from https://calhoun.nps.edu/handle/10945/1330

INITIAL DISTRIBUTION LIST

1. Defense Technical Information Center
   Ft. Belvoir, Virginia

2. Dudley Knox Library
   Naval Postgraduate School
   Monterey, California