Developing a Scaled Performance Evaluation Measurement System (SPEMS)

Garrett A Loeffelman, Capt, USMC, Quinn Kennedy, PhD, Glenn A Hodges, PhD, LTC, USA
Naval Postgraduate School
Monterey, CA
garrett.loeffelman@usmc.mil, <mqkenned/Gahodges1>@nps.edu

ABSTRACT

Training developers currently 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 and missions. The performance evaluation checklist (PECL) is the current binary standard that is used across all tasks to evaluate performance. SPEMS was also defined to be task agnostic but provides a quantitative alternative for evaluating performance. SPEMS leverages the thorough training and readiness task construct while adding the layer of numerical granularity that is necessary to reliably measure performance. We used the buddy rush task as a use case to test SPEMS and compare it to the current "Go/No Go" PECL. We developed SPEMS in three steps: we convened two focus groups to establish a 5-level behaviorally anchored rating scale; confirmed SPEMS reliability during three SME virtual video analysis focus groups; 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 from subject matter expert survey feedback as the most accurate objective measure of buddy rush performance. As a result, SPEMS/PECL were tested against exposure using linear regression. Fifty-two trainees were evaluated by a PECL and SPEMS evaluator simultaneously during three runs of the buddy rush task. The results showed that SPEMS has a moderate, negative, linear relationship with exposure at an $R^2 = 0.41/0.40$ compared to PECL at an $R^2 = 0.03/0.2$. The results also demonstrated that SPEMS is a more consistently reliable performance evaluation tool than PECL. We conclude that SPEMS scores are significantly related to percent exposure and have more predictive strength then PECL scores. These findings demonstrate a verifiable, repeatable, and reliable method for measuring military task performance across training solutions.

ABOUT THE AUTHORS

Capt Garrett A. Loeffelman is an infantry officer and recent Naval Postgraduate School (NPS) Modeling, Virtual Environments, and Simulation (MOVES) graduate. Capt Loeffelman is currently working as a modeling and simulation officer at Training and Education Command (TECOM). His research includes training effectiveness, return on investment analysis, and live, virtual, constructive (LVC) integration.

Dr. Quinn Kennedy is a research psychologist at the Naval Postgraduate School who conducts behavioral science research focused on human performance training and assessment.

LTC Glenn A. Hodges PhD is an Assistant Professor and Deputy Director for the Modeling, Virtual Environments, and Simulation Institute at the Naval Postgraduate School. His research includes human behavior representation, training effectiveness, and conceptual modeling.

Developing a Scaled Performance Evaluation Measurement System (SPEMS)

Garrett A Loeffelman, Capt, USMC, Quinn Kennedy, PhD, Glenn A Hodges, PhD, LTC, USA
Naval Postgraduate School
Monterey, CA
garrett.loeffelman@usmc.mil, <mqkenned/Gahodges1>@nps.edu

INTRODUCTION

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 (PECLs) that inform Marines if they are trained in a task. However, these measures do not inform Marines of their level of proficiency. According to the Government Accountability Office Report 13-698 *Better Performance and Cost Data Needed to More Fully Assess Simulation-Based Efforts*, the USMC lacks adequate performance data to determine the benefits of integrating LVC simulation capabilities into its current training programs (Government Accountability Office, 2013).

The primary reason the USMC lacks adequate performance data is that the performance of military tasks is challenging to quantitatively measure. A measurement system that is capable of not only comparing a Marine's performance to their peers, but also a training system's ability to improve a Marine's performance of tasks is desired. Our work demonstrates how a scaled performance evaluation measurement system (SPEMS) could be employed to show the benefits of training in different environments. The buddy rush task was selected as a use-case that demonstrates SPEMS applicability to small scale collective infantry tasks, but it is recommended that continued testing be conducted to generalize the findings of this proof of concept study across tasks and missions.

This research was addressed in three phases. Phase one consisted of defining the problem and exploring the background information to support a potential solution. Phase two consisted of two sets of pilot testing that centered on testing that SPEMS is a consistently reliable tool for evaluating task performance. Phase three was the conduct of an empirical evaluation of SPEMS and PECL in the operational environment. Researchers conducted a side-by-side comparison to determine which technique provided a more consistent, accurate, and predictive method for quantitatively evaluating performance. This phase concluded with the analyses to test our hypotheses and draw conclusions. Next steps include future experimentation aimed at generalization, empirically testing a proposed training system, and developing a return on investment model for system integration. These findings demonstrate a verifiable, repeatable, and reliable potential solution to the problem of measuring military task performance.

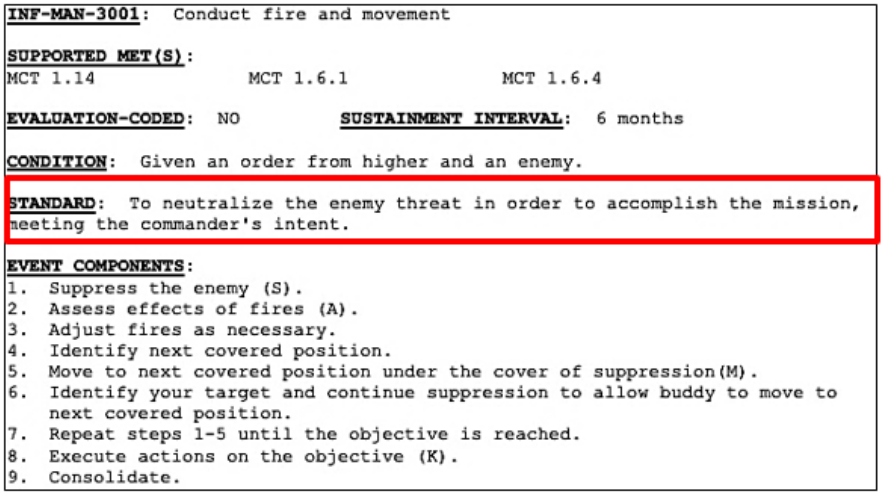
PHASE ONE: THE PROBLEM STATEMENT AND BACKGROUND RESEARCH

The Problem

The Marine Corps needs the ability to quantitatively measure the performance and readiness benefits of integrating expensive synthetic training environments into current training. These training systems are designed to support the performance of military tasks. The problem is that there currently does not exist a quantitative, reliable, or consistent way to measure the performance of these tasks. All USMC tasks are built on the individual training standard (ITS) that provides the framework for how the USMC evaluates performance. These tasks currently reside in the training and readiness (T&R) manuals that establish a hierarchical breakdown by military occupational specialty (MOS) of all tasks a Marine can train (Marine Corps, 2011). Individual performance is based on evaluating these tasks using a "Go/No Go" binary PECL system. This method of evaluating tasks introduces error and bias because evaluators are forced to reduce their understanding of performance to a binary judgement. Even though evaluators are trained during a multi-phased training and education process to identify what optimal performance looks like in each task, exhaustive qualitative feedback in the form of the after-action-review is reduced to this binary determination. This reduction limits how units can compare their performance in tasks to their peers or to the organizational average. Furthermore,

two systems that purportedly can 'train' the same tasks are believed to be equivalent because their performance benefits have not been quantified. We can more fully understand the dysfunctionality related to the current method of evaluating Marine task performance by examining two examples of T&R tasks that show the impact of ambiguous standards.

The task we examined is "INF-MAN-3001: Conduct Fire and Movement" (see Figure 1), commonly referred to as the "buddy rush". 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 (called event components in Figure 1), which must be executed to the standard. Our work 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.



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)

According to the USMC, the standard "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). One 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 measure of performance (MOP) for evaluating performance. Conversely, the standard in Figure 1 provides no quantitative metrics or MOPs and does not adequately meet the Corps' requirement that it "indicates the basis for judging the effectiveness of the performance…by…identif[ying] the proficiency level expected" (Marine Corps, 2011, p. 4-3). This standard instead allows for a wide degree of acceptable performance which can more accurately be described as a measure of effectiveness (MOE). Judgement, built on experience, must be utilized in order to determine proficiency using these MOEs (Loeffelman, 2019). However, this begs the question, is it possible for one buddy pair to complete a task more effectively than another buddy pair if it is based on judgement? Is it measurable in any sort of quantifiable way? Douglas Hubbard (2014) states, "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). Due to the wide breadth of tasks that need to be evaluated, we need a measurement system that is capable of plugging into the T&R framework and affords evaluators the opportunity to evaluate performance quantitatively and provide directed feedback. Quantifiable MOEs for all tasks could provide the performance data necessary to determine the effectiveness of training programs and systems. The next section consists of a literature review that guided the development of such a system, SPEMS.

Background

We conducted a literature review focused on three main areas: how the Marine Corps conducts and evaluates training, how to guard against bias and improve inter-rater reliability, and what training evaluation techniques currently exist in other fields. Focusing on how training is conducted provides the framework within which SPEMS must work. This

context is essential for adoption. Using subject matter expert (SME) raters is necessary for evaluating complex tasks; however, there are known issues associated with human involvement in rating. We address these issues by illustrating how baselining, anchoring, and training improve reliability well beyond acceptable levels. Finally, other fields have employed training evaluation techniques that have a track record of success. We dive into Kirkpatrick's four levels of training evaluation as well as discuss why behaviorally anchored ratings scales (BARS) provide a reliable, consistent, and accurate task evaluation framework.

How the Marine Corps Conducts Training

Military training is the foundation that prepares individuals to risk their lives by conducting physically and mentally challenging missions (Fletcher & Chatelier, 2000). The consequences of combat motivate 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. In this paper, we focus primarily on the USMC T&R manual.

Every job in the USMC is known as an MOS and consists of a set of tasks or skills required to complete that job (Marine Corps, 2004). The system's approach to training formalized these skills into the ITS. SMEs across the Marine Corps broke down each MOS into the skills and abilities required to perform a specific job. Each ITS includes six parts: the task, condition, standard, performance steps, administrative instructions, and references (Marine Corps, 2004). According to The Unit Training Management Guide, 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, 2016b, 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, 2016b, p. 4–3). ITSs became the foundation of how every single Marine is trained through the development of the T&R manual.

The USMC currently uses a binary mechanism for evaluating if tasks were trained, or untrained, depending on whether or not the task was performed to a given standard (Marine Corps, 2016b). As noted previously, a majority of the tasks are not easy to quantify; as a result, meeting the standard is done through the observation of completing each performance step. If the performance steps were completed in accordance with the standard, then a Marine is deemed to be trained in the task. The Marine Corps utilizes a "crawl, walk, run" to train each task that entails education, deliberate practice, live-fire qualification, and the after-action review (Trabun, 2007). This method is critical for layering tasks in what is known as a "building block approach to training," wherein Marines practice and are evaluated on progressively more complex tasks to become ready for combat (Marine Corps, 2011). As Marines move from initial training, to formal schooling, and onto the operational environment, skills compound to the battalion and regimental level. Unit performance is measured by aggregating binary task evaluations at lower levels into a higher-level computed readiness percentage (CRP). Wong et. al noted that averaging diverse performance standards typically confounds the accuracy of the CRP (Wong, Gerras, & Barracks, 2015). *Because commanders are evaluated according to the CRP, the flawed task evaluation system ultimately leads to a misrepresentation of a commander’s overall performance.* T&R tasks are the foundation to the entire training system; a flawed method for evaluating T&R task performance could cause inefficiencies and the misrepresentation of a unit's combat readiness.

These opportunities for error in the current training framework led to a review of current military training assessment conducted by Boldovici, Bessemer, and Bolton (2002). They outlined a number of principles that should be used to properly evaluate training and training systems. The three essential principles are: reliability, validity, and generality. The development of SPEMS focused on validity and reliability and offers a path to generalization for future projects. The Marine Corps does a good job of baselining and training evaluators; however, current field trail scores depend on evaluators boiling down performance to an unreliable binary measurement. This aspect of training evaluation could be improved by developing a quantitative anchoring system. We developed SPEMS to address this gap and tested its reliability and validity by conducting inter-rater virtual video analysis and operational testing (See Page 6).

Military Rater Reliability and Resistance to Bias

Evaluating complex and ambiguous situations using highly calibrated measurement instruments is often next to impossible (Hubbard, 2014). Tape measures can be used to measure objective qualities like dimensions, but complex behavioral situations require the use of human beings. Unfortunately, the human mind introduces a number of biases that influence the reliability of human judgement. Two common 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). However, Wigdor and Green's *Performance Assessment in the Workplace: Volume 1* demonstrated that military raters are extremely resistant to these common pitfalls due to significant baselining, standardization, and experience (Wigdor & Green, 1991). As stated by Boldovici, as long as rating reliability and validity are both measured and controlled, human raters are a viable option for measuring task proficiency (Boldovici et al., 2002). Wigdor and Green concluded that military jobs can be accurately modeled as a collection of tasks as organized in the T&R manual; these task therefore can be measured to determine job performance (Wigdor & Green, 1991).

Wigdor and Green (1991) asked 150 infantrymen 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. 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. All tasks were comprised of multiple performance steps that were evaluated according to the standard binary PECL in the T&R task's predefined order (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.

A Review of Training Evaluation Methods and Techniques

Kirkpatrick outlined his four levels of training effectiveness evaluation in 1959 that 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, but the military often fails to advance past the first level (Fletcher & Chatelier, 2000). The military's inability to take full advantage of this training effectiveness evaluation methodology is largely due to a lack of quantifiable performance standards.

BARS provide measures that are defined as “performance dimensions and scale values in behavioral terms” (Schwab, Heneman Herbert G., & DeCotiis, 1975, p. 550). 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 trainee (Schwab et al., 1975). Raters are able to act as observers, and the inferential requirements to judge task performance are left to those who create the BARS. BARS are developed using an iterative process where subject matter experts provide the critical incidents associated with the task, group these incidents by expertise level, and rate the incidents associated with each expertise level on their ability to represent the level of performance (Schwab et al., 1975). This process is repeated, and agreement is measured to determine what anchors will be used in the final rating scale. In the case of the T&R task, each task establishes performance steps that serve as the critical incidents necessary for completing the overall task (Marine Corps, 2016a). However, simply asking raters to determine how well a trainee 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 performance could redefine the value of a system by how it fosters quantitative improvements of a Marine’s proficiency in a task.

Performance Data's Role in Determining a Training System's Value: A Proof of Concept

Performance evaluation and analytic ratings provide opinions on how a training system is able to support training. These ratings should be applied to Marines' collective and individual performance of 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 SMEs to conduct a total system evaluation. The problem with current rating systems is that they measure a system’s ability to support the training of specific tasks as an absolute, but do not verify how well the system supports performance improvements. “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) 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 reason Dunne et al. was able to conduct this proof of concept study was because each M1A1 exercise is uniquely composed of 10 collective tasks scored within a range of 0-100. This distinctive quantitative characteristic offered the critical finding of the study; 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). 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 proficiency. Dunne's study provides an example of how capturing task proficiency using SPEMS could be leveraged, but training in the AGTS would have to be empirically compared to current training methods to determine the *training value* of the simulator.

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). Current training effectiveness evaluations use analytic SME driven processes such as the systematic team assessment of readiness training process (Dunne et al., 2017). These processes are used to determine what T&R tasks the system is capable of supporting. However, these processes do not answer the question of how well the system supports training the given tasks. Missing from these methodologies are quantitative ratings of individual and team performance that allow for performance gains to be identified. 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. In this type of study, SPEMS could provide a comparative performance measurement that demonstrates how training in each respective environment benefits a Marine's task proficiency (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. 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 can link cost analyses with performance improvements to ensure systems are acquired on the basis of training value.

PHASE TWO: PILOT TESTING

Introduction to Phases Two and Three

The empirical research portion of this study was conducted during phases two and three. Phase two 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 buddy rush task. Phase three consisted of an experiment that was conducted at the School of Infantry (SOI) West, Camp Pendleton, to collect data on the ability of current (PECL) and proposed (SPEMS) performance evaluation systems to capture trainee performance. The experiment manipulated one factor, the performance measurement technique, measured at two levels: current PECL scoring, and SPEMS. A paired design was used, such that a control evaluator and an experimental evaluator each evaluated the same trainee at the same time. A detailed description of the research methodology and results is described 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.

Two Pilot Studies for Developing SPEMS

Phase two entailed convening two different types of pilot tests that leaned on SME judgement to define and develop SPEMS. The first pilot test conducted a card sorting task that defined the SPEMS' BARS based on agreement, and the second pilot test involved SME evaluation of 15 virtual videos to determine the reliability of SPEMS. These pilot tests were conducted iteratively, and SPEMS was refined throughout the process. Phase two concluded with a developed, reliable SPEMS and validated measures of performance for phase three.

Pilot Test 1 - Methodology

The first pilot test consisted of two separate and distinct parts. During the first part, a subset of seven focus group members, split into three teams of two members, was asked to group and rank behavioral anchors according to their performance level to define SPEMS' BARS. The teams were provided a list of terms (anchors) that could be used to describe the performance of a military task between levels 1 (worst) and 5 (best). Each team was given 30 minutes to sort and rank the given terms, as well as add any terms they felt would help accurately capture performance at the given level (Loeffelman, 2019). Results were recorded, and the additional terms were added to the list for the second iteration of the task.

For the second iteration of the task, 33 anchors were included in the list of terms to be sorted by the focus groups. The same seven participants were asked to repeat the card sorting task individually; however, they could not add any additional terms. Again, each anchor was grouped and ranked within the five levels being told that level 1 indicated the worst performance and level 5 indicated the best performance. Rankings were determined by the participants based on 1 being the anchor that would most accurately indicate performance at that level. Anchors with higher rankings indicated they less accurately captured performance at that level. This concluded Pilot Study 1.

Pilot Test 1 - Results

Agreement was determined by analyzing what level participants placed each anchor in and the ranking it received in that level. This was done by counting how many of the seven participants placed the same anchor in the same level, and what the mean ranking of each anchor was in that level. For example, if all seven participants placed the same anchor in the same level it would have 7/7 agreement or 100%. Anchors were retained based on their percentage agreement and ranking. Table 1 illustrates the results of the card-sorting task.

Table 1. Card Sorting Pilot Test 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 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 BARS for SPEMS. A close approximation of these initial BARS can be seen under the scoring section of Figure 3 in their final state.

Pilot Test 2 - Methodology

The second pilot study consisted of the creation of 15 video vignettes of simulated buddy rushes that were evaluated during three separate focus groups. Infantry officer SMEs used SPEMS to evaluate each buddy rush video as if it were live. We trained research team members on proper buddy rush procedures and made videos capturing them conducting the buddy rushes in Virtual Battle Space 3 (VBS3). The team's actions were captured using standard video playback software to allow VBS3 simulations to be turned into test videos. Fifteen videos were developed to demonstrate a buddy pair conducting INF-MAN-3001: Conduct fire and movement (buddy rush) at various levels of proficiency. A screen shot of one of the videos can be seen in Figure 2. Once the test videos were created, they were further refined to most closely mirror realistic behaviors at various levels of proficiency.

Once the 15 videos were developed, we convened three focus groups of a total of ten infantry officers to evaluate each buddy rush using SPEMS. Infantry officers were specifically chosen for the focus groups to leverage their training,

baselining, and experience in evaluating the buddy rush task. This assumption is a realistic and relevant reality of training evaluators that is supported 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 time in service of seven years. All participants agreed they were qualified in the task and felt more than 90% familiar.

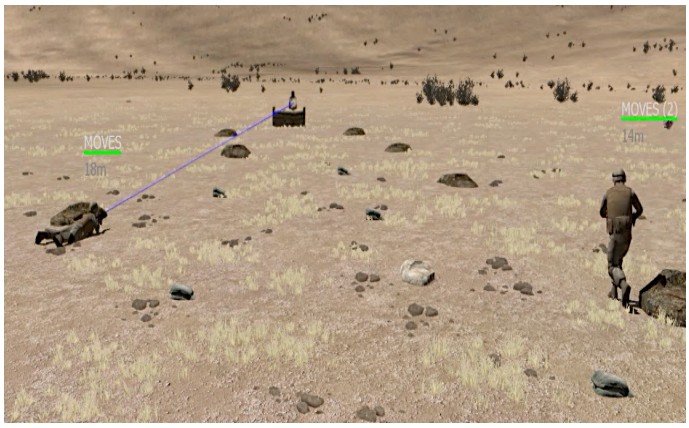


Figure 2. Virtual Depiction of "INF-MAN-3001: Conduct Fire and Movement"

During the conduct of each focus group, researchers elicited SPEMS scores from each participant following each video. Each participant was asked to watch each video for an unlimited amount of time and evaluate each performance step of the buddy rush using their SPEMS sheet. Scores were anchored by BARS, and the total task score was calculated by averaging the individual performance step scores for the whole task. Once all members scored a video, all SPEMS sheets were collected and participants' scores were reviewed for discrepancies greater than 2 levels. If a discrepancy occurred, discussions ensued to refine SPEMS prior to the next focus group. 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.

Pilot Test 2 - Results

For Pilot Study 2, inter-rater reliability was measured by calculating the Cronbach's alpha for each focus group, which provides evidence that all raters are evaluating the same underlying concept, the buddy rush, with similar results. All three focus groups (consisting of three to four participants) demonstrated high degrees of inter-rater reliability on the overall performance scores (Cronbach's Alphas ranging from 0.93 to 0.98). Overall performance scores were calculated as the average SPEMS rating across all performance steps. This high degree of reliability across all focus groups validated the reliability of SPEMS prior to operational experimentation resulting in the SPEMS scoring sheet in Figure 3. Furthermore, 100% of participants across all focus groups indicated they felt SPEMS was more effective than the PECL evaluation method with an ease of use score of 8.8/10 (s = 0.76, 95% CI = (8.33, 9.26)) and an effectiveness score of 9.1/10 (s = 0.69, 95% CI = (8.67, 9.53)).

Trainee Name: _____

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 enemy (S). 1 2 3 4 5

Figure 3. Final SPEMS Scoring Sheet used during Experimentation

Researchers also required an objective MOP for the buddy rush task to test the strength of SPEMS scores' relationship with task performance. The buddy rush MOP was determined through an iterative process across the three focus groups through survey responses and discussion. Through this process, the amount of time a buddy pair is exposed was unanimously selected as the most accurate buddy rush MOP. 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. This proxy measurement was validated by further soliciting SME feedback.

PHASE THREE: LIVE EXPERIMENTATION AND RESULTS

Phase three consisted of the experiment conducted at 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.

Methodology

Trainees at SOI West conducted 4 days of deliberate blank fire and live fire practice of INF-MAN-3001: Conduct Fire and Movement (the buddy rush) as part of their standard training practices while two independent SPEMS/PECL evaluators rated their performance. Following recruitment and consent, no adaptations were made to the trainees standard training schedule. Instructors from SOI's combat instructor pool volunteered, consented, and were assigned to be either PECL or SPEMS evaluators randomly. PECL evaluators verbally verified their qualification in the task of evaluating the buddy rush and were promptly dismissed to maintain their blindness to SPEMS. SPEMS evaluators were trained on SPEMS by watching a distributed set of 5 buddy rush videos at various proficiency levels. Evaluators were told the average infantry officer rating of each video for baselining.

During the conduct of each run, two buddy pairs were each independently evaluated by one SPEMS and one PECL evaluator. Twenty-six buddy pairs (52 Marines) were evaluated by both a SPEMS and PECL evaluator over the course of three runs. Simultaneously, the research staff counted the total number of rushes and the total number of exposed rushes for each buddy pair. An exposed rush is defined as one buddy advancing towards the target without suppression. "Without suppression" was determined by observing if the pop-up target was up at the same time that a buddy is moving towards the target. The percent exposure was calculated using the following equation.

$$\%\, Exposure = \frac{Number\ of\ Exposed\ Rushes}{Total\ Number\ of\ Rushes} \qquad (1)$$

The experiment was completed after the last buddy pair was evaluated and measured on the third run.

Preliminary Results

The distribution of scores for overall performance and by performance step by evaluation type (SPEMS / PECL) are described in Table 4. The PECL overall performance 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 (shown in Figure 1). 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 2. 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 for percent exposure appeared approximately normal with no statistically significant difference in the average percent exposure between run 2 (61%) and run 3 (54%) (t(25) = 1.71, p = 0.09). Distributions for both runs percent exposure visually appear to be approximately normal. The normality of these distributions more closely resembles the more normal distribution of SPEMS scores, which suggests there may be a relationship between SPEMS and percent exposure. In the next section, we formally tested this potential relationship.

Results

The primary results are focused on testing the following hypotheses. Overall SPEMS and PECL scores were used, and these overall scores were calculated as the average of all performance step scores.

Hypothesis 1 Testing

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

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

Run 2: Percent Exposure = 1.216 - 0.162*SPEMS Score
Run 3: Percent Exposure = 1.480 - 0.268*SPEMS Score

The assumptions and conditions for linear regression were met. Figure 4 shows a statistically significant moderate negative, linear relationship between SPEMS scores and percent exposure across the two Runs (p-values <.0006): 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. R^2 values consistently explain approximately 40% of the variability in percent exposure. *Therefore, we reject our null hypothesis and conclude that there is a negative relationship between SPEMS scores and percent exposure.*

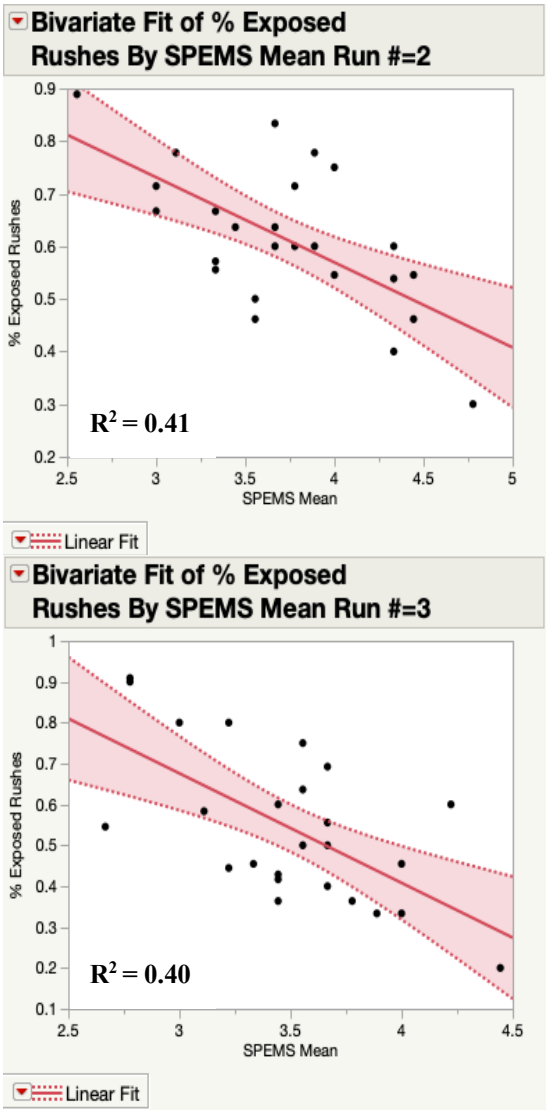


Figure 4. Linear Regression Results Testing Fit of Mean SPEMS Scores and Percent Exposure. Both Runs Demonstrate Moderate Fit.

Hypothesis 2 Testing

H_0: There is no difference in the predictive strength between overall SPEMS scores and overall PECL scores 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.

We tested the second hypothesis by conducting linear regressions between PECL scores and percent exposure and comparing the results to those using SPEMS scores as the predictor variable. We first checked the assumptions and conditions and found that Run 2 PECL data did not meet the linearity, equal variance, or independence assumptions. Run 3 did meet the assumptions and conditions, but there are concerns regarding the equal variance assumption. Results should be interpreted with caution. The PECL linear regression model for Run 3 is shown in the following equation:

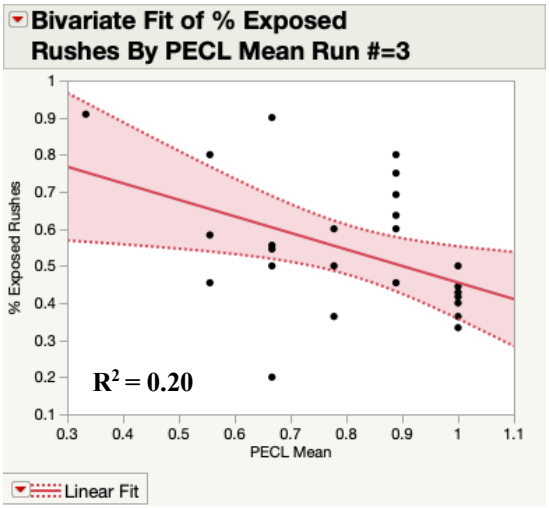


Figure 5. Linear Regression Results Testing Fit of Mean PECL Scores and Percent Exposure. Run 3 demonstrates poor fit.

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

Run 3 has an R^2 of 0.21 demonstrating a weak, negative, linear relationship between PECL scores and percent exposure (see Figure 5). Although the slope results are significant ($t(25) = -2.49$, $p = .020$) the R^2 value indicate that PECL scores only explain 21% of the variability in percent exposure. Because the PECL data does not adequately meet the assumptions and conditions, this statistical result should be viewed with caution. If we revisit hypothesis 2 and compare these results to those from SPEMS scores, we see that the SPEMS model has greater predictive strength based on meeting the assumptions and conditions, having more significant slope estimates, and much larger R^2 values. *We reject the null hypothesis and conclude that SPEMS scores have more predictive strength than PECL scores.* Below, we discuss two points of interest that emerged from the data collection.

Other Findings

First, when conducting the card-sorting task, 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, the anchor "No Go" was placed in the 1st level for proficiency by 86% of participants with a mean rank of 5.00 (s = 2.28). This finding could indicate that "No Go" is commonly associated with a level 1 out of 5 or as low as 20%. As a result, evaluators using the current PECL technique could perceive "Go" as achieving proficiency in a task greater than a 1 out of 5 on a performance step. The Marine Corps anecdotally utilizes 80% as its passing standard, and this finding could indicate a difference in what is considered passing by the organization versus the evaluator.

A second point concerns the moderate strength of the relationship between SPEMS scores and percent exposure evidenced by an $R^2 = 0.41/0.40$. This relationship may have been weakened by error associated with the targets. Targets respond to accurate suppression by sensing the impact of a round on their target face; however, as the target face is degraded, so is the sensing ability of the target. If we assume this error correctly and remove data points corresponding to where percentage exposure is significantly higher than the SPEMS score (indicating possible target issues) we see R^2 values of 0.64 (Run 2) and 0.60 (Run 3). There was no way to verify a target malfunction during these operational experiments; therefore, these results were not included. More testing should be conducted that ensures the reliability of targets and more accurately measure the relationship between SPEMS and this MOP.

CONCLUSIONS & RECOMMENDATIONS

This work demonstrated a verifiable, repeatable, and reliable method for measuring military task performance across all training solutions. We did not find PECLs to be an effective or reliable performance evaluation method. In contrast SPEMS provides a tested potential alternative for quantitatively evaluating performance. Standard card-sorting

techniques were used to develop BARS for SPEMS; the reliability of SPEMS was validated through virtual video analysis; and the predictive strength of SPEMS was empirically tested through operational testing.

SPEMS scores have more predictive strength than PECL scores on task performance, illustrating a more viable method for evaluating Marine Corps tasks. SPEMS has the potential to impact the training and acquisitions domain by providing a reliable, consistent, and quantitative performance measurement system. 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. SPEMS does this by providing valuable insights into what performance steps are most affecting overall task performance. In contrast to the inconsistent PECL data, performance data can allow the training staff to focus future deliberate practice on these highlighted areas of weaker performance.

In the acquisition domain, SPEMS provides the quantitative data for evaluating how a training system supports the improvement of a Marine's performance. Currently, developers are able to determine the life-cycle costs of a system, estimate the costs the system avoids through simulation, and decide which T&R skills can be trained using a TEE. The problem with this data is that it does not take into consideration the actual performance improvement gained by the users. Training systems like the Indoor Synthetic Marksmanship Trainer 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 the system supports training the task (Loeffelman, 2019). We recommend SPEMS data be used to conduct side-by-side comparisons of training solutions to select methodologies that maximize proficiency and optimize cost avoidance.

This work demonstrates an important framework for how further investigation into SPEMS' ability to evaluate performance should be conducted. To continue developing SPEMS to meet the above recommendations, we see the following four steps as necessary. 1. Generalize SPEMS usability across tasks and missions by conducting multiple proof-of-concept experiments in different tasks to demonstrate generalizability. 2. Empirically test the integration of a proposed training system in a side-by-side experiment with current training programs to demonstrate proficiency and cost avoidance advantages. 3. Combine performance data with life-cycle cost and cost avoidance data to determine the return on investing in proposed training programs. 4. Validate this new model by monitoring integrated training solutions throughout their life cycle to test the accuracy of estimated performance/cost models (Loeffelman, 2019). The development of SPEMS serves as a verifiable, repeatable, and reliable potential solution to the problem of measuring the benefits of integrating synthetic training to improve Marine Performance across tasks and missions.

ACKNOWLEDGEMENTS

This research was funded by the Office of Naval Research Code 34. We would like to thank the Synthetic Training Integration Management (STIM) at Training and Education Command (TECOM), the leadership at the School of Infantry (West) Infantry Training Battalion, and the NPS staff for helping us tackle this important problem.

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/

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

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

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.

Loeffelman, G. (2019). *Developing a scaled performance evaluation measurement system to evaluate Marine performance.* (Master's Thesis). Retrieved from https://calhoun.nps.edu

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. (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. (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. (2016b). *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

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

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

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

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