NAVAL POSTGRADUATE SCHOOL

MONTEREY, CALIFORNIA

THESIS

MODELING AND SIMULATION WARGAMING TOOL FOR NAVY STAFF OFFICER TRAINING

by

Daniel L. Cain

June 2019

Thesis Advisor:	Jeffrey A. Appleget
Co-Advisor:	Perry L. McDowell

Research for this thesis was performed at the MOVES Institute.

Approved for public release. Distribution is unlimited.

THIS PAGE INTENTIONALLY LEFT BLANK

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

Public reporting burden for this collection of information is estimated to average 1 hour per response, including the time for reviewing instruction, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing the collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden, to Washington headquarters Services, Directorate for Information Operations and Reports, 1215 Jefferson Davis Highway, Suite 1204, Arlington, VA 22202-4302, and to the Office of Management and Budget, Paperwork Reduction Project (0704-0188) Washington, DC 20503.

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

4. TITLE AND SUBTITLE MODELING AND SIMULATION WARGAMING TOOL FOR NAVY STAFF OFFICER TRAINING	**5. FUNDING NUMBERS**
6. AUTHOR(S) Daniel L. Cain	
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)

We investigated utilizing serious games to train officers on Navy operational staffs, such as carrier strike groups (CSGs), destroyer squadrons (DESRONs), and amphibious squadrons (PHIBRONs). Such staffs are composed of officers from different warfare communities, such as aviation, submarines, and surface warfare. Most have not served on such a staff before and have limited familiarity with the inner workings and responsibilities of their command. We reviewed the current standards of training and serious games usage, and designed an experiment to determine whether serious games could provide a statistically significant improvement in training transfer for deployable staff officers compared to traditional methods of training. The experiment group was composed of West-Coast watchstanders. They played two different scenarios on two separate gaming applications for a total of four sessions. We compared performances between the experimental group and control group using a pre-test and a post-test given after the training. We also conducted another test one month later to see if a difference existed in long-term retention. The control group was enrolled in the joint maritime tactics course, using classroom lectures administered by Tactical Training Group, Atlantic. Small sample size merited nonparametric statistics usage, which increased the difficulty of obtaining significant results. Only one test produced a significant outcome, but the results feed into demand for future work.

14. SUBJECT TERMS modeling and simulation, wargaming, serious games	**15. NUMBER OF PAGES** 89
	16. PRICE CODE

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

THIS PAGE INTENTIONALLY LEFT BLANK

MODELING AND SIMULATION WARGAMING TOOL FOR NAVY STAFF OFFICER TRAINING

Daniel L. Cain
Commander, United States Navy
BS, San Diego State University, 2000
MBA, Webster University, 2009

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: Jeffrey A. Appleget
Advisor

Perry L. McDowell
Co-Advisor

Peter J. Denning
Chair, Department of Computer Science

THIS PAGE INTENTIONALLY LEFT BLANK

ABSTRACT

We investigated utilizing serious games to train officers on Navy operational staffs, such as carrier strike groups (CSGs), destroyer squadrons (DESRONs), and amphibious squadrons (PHIBRONs). Such staffs are composed of officers from different warfare communities, such as aviation, submarines, and surface warfare. Most have not served on such a staff before and have limited familiarity with the inner workings and responsibilities of their command. We reviewed the current standards of training and serious games usage, and designed an experiment to determine whether serious games could provide a statistically significant improvement in training transfer for deployable staff officers compared to traditional methods of training. The experiment group was composed of West-Coast watchstanders. They played two different scenarios on two separate gaming applications for a total of four sessions. We compared performances between the experimental group and control group using a pre-test and a post-test given after the training. We also conducted another test one month later to see if a difference existed in long-term retention. The control group was enrolled in the joint maritime tactics course, using classroom lectures administered by Tactical Training Group, Atlantic. Small sample size merited nonparametric statistics usage, which increased the difficulty of obtaining significant results. Only one test produced a significant outcome, but the results feed into demand for future work.

THIS PAGE INTENTIONALLY LEFT BLANK

TABLE OF CONTENTS

LIST OF FIGURES

THIS PAGE INTENTIONALLY LEFT BLANK

LIST OF TABLES

THIS PAGE INTENTIONALLY LEFT BLANK

LIST OF ACRONYMS AND ABBREVIATIONS

CCSG	commander, carrier strike group
CWO	chief warrant officer
DoD	Department of Defense
FST	fleet synthetic training
IRB	institutional review board
JMTC	joint maritime tactics course
LITMUS	littoral combat ship integrated toolkit for mission engineering using simulation
MET	mission essential task
NDM	naturalistic decision making
NPS	Naval Postgraduate School
OFRP	Optimized Fleet Response Plan
OLW	operational level of war
ROE	rules of engagement
RPD	recognition-primed decision
SG	serious game
SOH	Strait of Hormuz
SME	subject matter expert
SWO	surface warfare officer
TEE	training effectiveness evaluation
TTP	tactics, techniques, and procedures

THIS PAGE INTENTIONALLY LEFT BLANK

ACKNOWLEDGEMENTS

I have so many people to thank. Without the support of those below and many others, I could not have completed this work.

First, I want to start with my advisors, COL Jeffrey Appleget (Ret.) and Perry McDowell. They shared their knowledge and experience that helped scope my thesis. Every challenge outside of my control was handled with professionalism. I would choose both of you again in a second.

Peter Squire and ONR—thank you for believing in my project and providing funding.

jCORE team—thank you for hosting Perry and me to really understand how you could help. Your selfless effort was commendable!

LITMUS team—thank you for taking time to participate in my thesis experiment. Zach, Clint, and Ben did an amazing job!

Rabia Khan—your devotion to learning the serious games and being flexible was truly impressive. I cannot tell enough people how much of a positive impact you had! I had a great time hanging out playing the games with you. You are a fun, interesting person.

Terry Norbraten—you rogered up to help out when I really needed it. My experiment turned out great because you were the jack of all trades by flawlessly executing whatever I needed.

Matt Timmerman—thank you for taking time to be my surrogate on the East Coast!

The carrier strike groups and their volunteers—thank you for devoting time that you all did not have to spare. I enjoyed working with you and learned a great deal from you!

Finally, the most important thank you is for my family. Without them, none of this is possible. I love you all so much. The last two years were great because of you.

THIS PAGE INTENTIONALLY LEFT BLANK

I. INTRODUCTION

A. RELEVANCE

U.S. Navy deployable staff officer billets are filled from a variety of communities like aviation, surface, subsurface, and even restricted line officers such as information warfare. This practice provides the flag officer or commodore with varying experiences and talents. Not all backgrounds offer a similar experience, even if they are from the same community. For instance, there are differences between officers stationed in forward deployed locations like Japan and those on the eastern seaboard, those with cruiser or destroyer (CRUDES) experience versus surface warfare officers (SWOs) who served on amphibious ships, or aviators flying expeditionary rotary aircraft compared with carrier air wing fixed wing fighters. It is impossible for an organization as large as the Navy to provide staff leadership with every unique skill and knowledge desired. There is not one ideal background, or one that is more successful than others with the exception of individual sustained superior performance. Each deployable staff officer should, like any job, have the expectations communicated and the tools to succeed.

Watchstanders prepare for deployment using he optimized fleet response plan (OFRP) construct. Officers reporting to a deployable staff are highly qualified on the platform they came from, but less familiar with other communities and associated systems' capabilities. During pre-deployment "work-ups" and throughout deployment they will have to do more than just interact with other platforms or staffs. The skills and knowledge of staff officers are only honed with training and experience in the job, and only a fraction of the previous years' experience applies. For instance, a top-performing SWO with years of experience in engineering and propulsion systems on DDG Flight IIA destroyers may not have the ability to communicate with "Whiskey" or "Zulu" for tasking a P-8A Poseidon intending to conduct intelligence, surveillance, and reconnaissance (ISR) during a strait transit.

B. CURRENT U.S. NAVY TRAINING

The Navy must train a diverse set of communities to a requisite level of readiness. To train and educate so many different members by rate and rank, they must employ a host of distinctive methods. Traditionally, the Navy would “train like you fight” on the actual system in a live training environment. There are other training tools available to augment and sometime replace live training.

1. Navy Training for Unrestricted Line

There are a host of training programs and curriculums relying solely upon the live environment. For example, to execute shooting a missile from an aircraft for training and readiness, many things must happen. There needs to be one or more mission-capable aircraft provided by the maintenance team, ordnance ordered and received from a combat aircraft loading area (CALA), a scheduled and approved area on a range, an ordnance load team, the qualified aircrew, and evaluation SME to determine if it was a valid expenditure. If one piece failed for any reason, the event would lose its usefulness, and must be scheduled.

Some emerging training options were helpful but limited, which meant that traditional training methods continued to be the standard. The Navy evolved its live training by including training ordnance like recoverable torpedoes and blank ammunition before embracing other technologies like virtual training in simulators. However, warship steaming and pilot flight time was still necessary to judge the crews ready for deployment. More days underway were also required than today since a legitimate alternative did not exist.

Today, technology improved and other training methods became possible. Training curriculums contain a mix of simulators, or ”sims,” and live training, with live still the preferred option when available. Previously, sims could only practice emergency procedures and novice cockpit procedures. Sims are becoming a more widely accepted and used resource since they consistently execute tactics, techniques, and procedures (TTPs), save time, money, and allows many repetitions safely while freeing up actual systems for operational use. Indeed, sims have expanded capabilities due to investments in fidelity upgrades over the last decade. Sims now there are more complex arrangements which link disparate platforms and locations together, which is a training concept called “LVC,” or live, virtual, constructive.

Unrestricted line communities train on more than just sims and live training. Navy members are familiar with and use initial classroom training, computer-based training (CBT), part-task trainers (PTT), to name a few. CBTs are mostly known as annual, passive training lessons accessible anywhere and anytime. PTTs train the user on a specific process or skillset. The 2018 keynote speaker of the Interservice/Industry Training, Simulation and Education Conference in Orlando, ADM Grady said, “We must apply modern training delivery models, understanding that everything isn’t best learned in a brick-and-mortar schoolhouse” (Lee, 2018). Though the Navy still uses classrooms to gather students, they employ lessons on tablets or practice on synthetic trainers like the multipurpose reconfigurable training system (MRTS).

2. Staff Training

Staff officers carry a heavy burden of required experience and a skillset including communication, decision-making, and comprehension of doctrine and leadership guidance. Navy deployable staffs use only a few training aids and some classroom training. First-time watchstanders usually attend classroom education for exposure to necessary staff officer skills. Through the OFRP, a staff will complete fleet synthetic training (FST). FST provides the staff an opportunity to practice as though deployed and executing a realistic scenario. Finally, the staff gets underway for a few weeks to train with the other staffs who they will deploy with later.

C. CAN A SERIOUS GAME HELP?

Deployable staffs need a variety of training options; two methods like FST and weeks-long live training may not meet all the individual needs. The OFRP outlines the training curriculum along with other requirements to be certified for deployment. Though extensive resources and planning go into executing a curriculum, there is always room for an outside perspective to analyze if their training can in any way be further optimized. One such optimization tool for training could be a serious game (SG).

Serious games have been used for decades to teach in different capacities than traditional teacher-student classroom methods. One of the first was *Oregon Trail*, an education game widely used in the 1980s. Increased computational power enabled a first-person shooter named *America's Army* (U.S. Army, 2002). This SG proved educational and motivational across a broader audience. Over the last decade, the SG niche has grown beyond academic curiosity with specific designs to accommodate adaptive learning techniques. The desire to integrate these possibilities has swelled alongside an ever-increasing appetite for electronic home entertainment systems and games.

Serious games are strong motivational tools compared to traditional learning means. "Fun" is a powerful force that leads to higher interest and curiosity levels (Iten & Petko, 2014). A group of players completing an SG makes them individually feel a sense of accomplishment, and perhaps desire continuing to learn without being prompted. The experience is unique even if everyone starts the SG at the same time. Conversely, in a classroom setting, if all students take a test and see they all received the same score then it minimizes the overall significance.

The time had arrived where properly designed games could inject more than just a respite from traditional learning methods. Just like the military's training community does not want to replace live training completely with virtual training, the gaming community is merely attempting to augment traditional learning with digital opportunities. The latest technology-savvy generation entering the military should have the most applicable learning techniques and training systems available to take advantage of their core skillsets.

D. SCOPE OF EFFORT

This thesis will investigate if serious games can provide better or more cost-effective training options for deployable staff officers than the training methods the Navy currently employs. The goal is to increase the knowledge for the following research questions.

1. What is the current standard, and are existing training methods meeting the standard?

2. What are the factors that a serious game should possess to effectively train Navy staffs?

3. Can serious games either replace some training that is currently being done to standard but is cost-ineffective, or fill a gap previously identified?

4. Do the serious games used provide a statistically significant difference in training transfer for deployable staff officers compared to traditional methods of training?

The first three research questions will be discussed in Chapter II. The last question uses the hypothesis claiming there was no difference in training between classroom instruction and playing serious games. Data is analyzed with survey data and a pretest/post-test/post-test experiment format. During the experiment, participants from Navy staffs played two different serious games over two days with two different scenarios. A control group was given classroom instruction. A statistical comparison of the experimental and control group using mixed-design ANOVA was used.

1. Omitted Areas of Study

An important point to emphasize is the experiment does not double as a training effectiveness evaluation (TEE). A TEE is a much more involved process, and considered a future work possibility.

Construction of a purposefully designed serious game is also outside the scope of this thesis. The time and resources required to properly design a serious game that meets a currently uncommunicated requirement is far greater of an undertaking and beyond this paper. Though this is a potentially important undertaking, it will be confined to the future work section.

2. Thesis Organization

Chapter II provides a background and foundation for understanding the thesis experiment. It provides detail about how a deployable staff trains for deployment, discusses the specifics of a serious game, and contains a literature review regarding pertinent topics. Chapter III covers the approach which includes the hypothesis, scenarios, methods for collecting data, and information on the participants. Chapter IV covers data compiled as well as the analysis and interpretation of this data. Chapter V finishes with a discussion of the experiment, some recommendations, and future work opportunities.

II. BACKGROUND

A. HOW A DEPLOYABLE STAFF GETS READY TO DEPLOY

Staff leadership has a different role than that of a squadron or ship commanding officer. The staff admiral or commodore is charged with planning documents, battle rhythm, training, and providing guidance to those operational units under their tactical control (TACON). The staff's guidance from higher leadership comes from instructions (Department of the Navy, 2013; Joint Chiefs of Staff, 2018; Department of the Navy, 2010) and the OFRP (U.S. Fleet Forces Command, U.S. Pacific Fleet Command, 2012; U.S. Fleet Forces Command, II Marine Expeditionary Forces, 2016), also known as the deployment cycle, details the process. The Navy deployment cycle is broken down into four phases. Scoping the OFRP to staffs is the next section, and will include a discussion on the main training and readiness aspects needed to be certified ready to deploy.

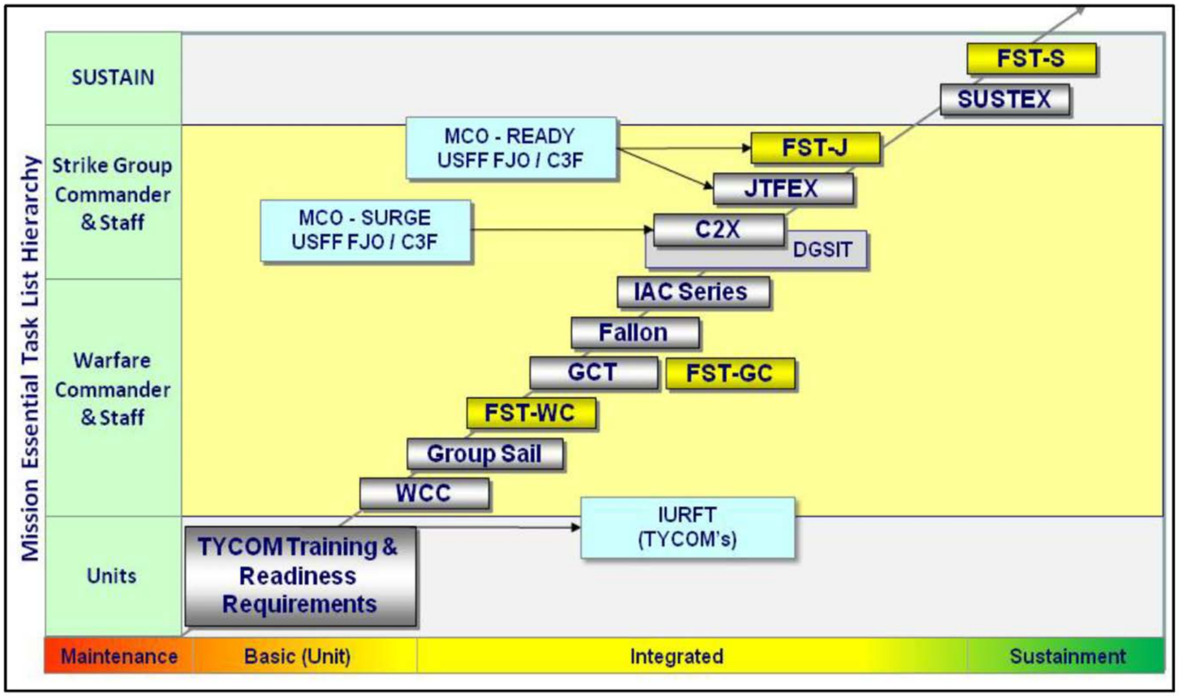
1. Navy Deployment Cycle

The Department of Defense (DoD) spends billions of dollars annually to train and increase readiness so units are prepared to deploy. Every task that the military trains to is tracible to a larger national defense strategy (DoD, 2018; President of the United States, 2017). The country's high-level strategy documents produce requirements that the military must fulfill, and these requirements create mission essential tasks (MET) for the units and individuals to meet. When each Navy unit has completed the assigned METs and an assessment unit concurs, a certification for deployment will be awarded. The goal of the staff is to achieve a certification that they are ready to deploy. The OFRP is the framework spelling out what is needed to achieve the certification. It breaks down into four distinct phases: maintenance, basic, integrated, and sustainment. A unit's training curriculum differs depending upon how close it is to deploying. The OFRP commences upon completion of the sustainment phase (when the unit is no longer either deployed or serving as a potential surge force).

The maintenance phase begins the entire deployment cycle. This is when units receive the bulk of major maintenance and upgrades. This is the ideal time to transition the support staff in order to send the new members to required schools. In this phase, units are not considered deployment-ready, and the Navy devotes minimal effort toward attaining high readiness levels at this juncture since most will not count toward the unit’s readiness score when certification is desired. Some training has a periodicity, meaning that after a few months it no longer factors into determining if the unit is ready to deploy. Most training during this phase is devoted to individual training

The next phase is the basic phase. The goal of the basic phase is completion of unit-level training (ULT). Members of the unit should complete as many individual requirements and internal unit needs as possible. Individual training expands to include team drills and practice, and exposure to who else is making the same deployment. Less complicated FST events are conducted here. Funding increases some in this phase to improve readiness but remains overall low since the actual deployment is still likely many months away. Readiness acts as a binary checklist of items achieved across the OFRP making it easy for leadership to measure completed and remaining tasks.

The integrated phase is the third and final pre-deployment phase. The training dimensions expand to the most rigorous and complex scenarios for the unit to include associated units deploying together. Integrated training, knowledge, and skills should reflect the maximized budgeting for readiness. By its completion, the organization should have completed in-port and at-sea training exercises like FST and composite training unit exercise (COMPTUEX). Figure 1 shows all the events for a strike group across the OFRP. FST-J, or the most complex “joint” version, is the final pre-deployment requirement after underway periods. This combines numerous units from multiple services practicing mission sets that are otherwise nearly impossible to coordinate considering schedules and costs.



Mission Essential Task List Hierarchy
SUSTAIN
Strike Group Commander & Staff
Warfare Commander & Staff
Units
FST-S
SUSTEX
MCO - READY USFF FJO / C3F
FST-J
JTFEX
MCO - SURGE USFF FJO / C3F
C2X
DGSIT
IAC Series
Fallon
GCT
FST-GC
FST-WC
Group Sail
WCC
IURFT (TYCOM's)
TYCOM Training & Readiness Requirements
Maintenance
Basic (Unit)
Integrated
Sustainment

Figure 1. Generic carrier strike group OFRP. Source: U.S. Fleet Forces Command, U.S. Pacific Fleet Command (2012).

The last phase is the sustainment phase. Sustainment can encompass a period prior to and the period between deployments. Sustainment funding is sufficient to maintain readiness till the next potential deployment. Here, a strike group staff attempts to maintain readiness as best as possible with the tools available like FST or even sending watchstanders to other staffs to sharpen skillsets. This is the glue to enable getting two deployments from the same training cycle, but any major personnel transfers can negatively impact the second deployment.

2. Staff Deployment Cycle

The OFRP, or work-up cycle, is an incredibly busy time for a deployable staff. The training aspects of the work-up cycle are not the only important items requiring attention. Like any deployable unit, there are tasks received from superiors and delegated to subordinates. Finding time to fit in all the levels of requirements and maintain a high standard of skill proficiency is similarly arduous. Knowledge and skill acquisition are perishable, especially for the officers serving for the first time on a staff. Staffs require a method to train officers which does not interfere with their other duties during the work-up cycles.

A staff watch officer serving for the first time normally attends a two-week course named Joint Maritime Tactics Course (JMTC) at tactical training group, Atlantic (TTGL) or Pacific (TTGP). This course exposes the new staff officer to primary warfare areas, the course of action (COA) process, battle rhythm for daily and weekly operations, understanding the interactions required between warfare commanders, and hands-on practice with a FST event. The staff officer may take other courses offered if more specific education is needed on the staff, the longest class taking over a month.

Another major training function offered throughout the OFRP for staff officers is FST. FSTs are typically dedicated staff officer training events that tie in virtual players and operators linked-in to contribute, as available and requested. FST can increase or decrease the complexity of the event based on the participants needs or proficiency, and can last from a day to a week. These adaptive traits mean FST can be employed in most any phase of the OFRP. FST is a cornerstone of the staff officer training curriculum since the only alternative is to embark the carrier and get underway for live training. Live training is saved for the integrated phase after conducting simulated “reps and sets” to improve watchstander competence and experience.

Dedicated courses for staff officers, FST and live underway training are the primary training options for watchstanders. This training is in addition to the all-Navy training requirements like general military training, annual CBTs, or physical fitness. The three primary options do not cover all training opportunities afforded to operators. There are no PTTs or specialized CBTs. This points to a gap to better prepare or sustain watchstander skillsets when preparing for deployment.

B. THE SCIENCE OF LEARNING AND ITS IMPACT ON GAMING

As technology has matured, there have been an increasing number of studies to better understand several pertinent questions, like whether an SG has tangible benefits. What field has the most effective integration tools? How does an SG compare to a traditional teacher-student pedagogical environment? To best understand how an SG transfer of training works, there are additional foundational studies worth noting. This section provides three parts: first, an overview of some relevant theories and best practices; next, is about SG performance assessment experiments and outcomes; finally, peer-reviewed SG studies that speak specifically to transfer of training.

A foundational piece for understanding the benefits of serious games is in the science of learning. One of the most widely accepted studies is Bloom's taxonomy (1956). Though cited for traditional classroom education, this makes sense as a starting point to field questions about SG learning qualities. This taxonomy was a baseline for creating surveys and test questions for participants in the experiment playing serious games and control group receiving classroom training. This established method used a hierarchical model for classification and learning objectives based on complexity and specificity. The six stages in order from lowest to highest are knowledge, comprehension, application, analysis, synthesis, and evaluation. Even though there are legitimate arguments since its acceptance from the field of education, it maintains its place as a relevant organizational tool from which to judge the process of learning.

Naturalistic decision making (NDM) is germane to learning sciences discussion since the military values experience when making decisions, especially since a military mistake can cost lives, precious equipment, or time. NDM was conceived in the 1980s, and is a framework to study how people come to their specific conclusions while inserted into complex, real-world circumstances. This can be a challenging method to glean significant outcomes from since there are potentially numerous variables dynamically changing during the test. From NDM came the recognition-primed decision (RPD) model, whose purpose was to try to explain how highly experienced people can quickly determine what information is critical to decision making while disregarding other seemingly pertinent data. NDM generally bins people's skill in a task from novice to mastery, based on speed of intuition. RPD was considered a factor for the experiment's players and how quickly they could properly employ TTPs.

The latest framework for instructional design to achieve consensus acceptance came from Dr. M. David Merrill (2002). His name for the process is first principles of instruction. Its intent established a method that withstands different fields of study and their associated learning needs. The central theory is that students achieve the best results in problem-centered issues. The first principles of instruction theory was considered as a refined Bloom's taxonomy for measuring volunteers scoring and retention. This concept is broken down into five principles. The first principle states any task promotes learning the best when solving real-world problems. Activation helps students learn more by recalling past experience to learn new skills, which is similar to scaffolding. Demonstration of new knowledge is another way that promotes learning by being shown vice just being told. Performing real-world tasks is also superior for learning instead of simple information passing from teacher to student. Finally, integrating the new knowledge into the learner's domain helps demonstrate understanding of the new learning topic.

Game-related learning assessment is wrought with questions still. Though there have been many studies, little is known about what serious game elements can impact outcomes for student learning (Van Staalduinen & De Freitas, 2011). There needs to be a balance between open-ended games where the student may not do what the instructional designers intended versus having to strict of a path that reduces player motivation. Recent research and experimentation continue to provide in-game assessment tools for designers, though they are still maturing (Dede, 2012). Researchers are conducting experiments with large sample sizes over many years with an emphasis on capturing valid methodology and datasets to better understand assessment (Mayer et al., 2014).

Just as assessment struggles to provide a definitive solution, showing a transfer of training is even more difficult:

> As such it is an important concept in determining training value. However, it can be difficult to determine what exactly is learned with respect to the (real) task or domain for which the training is intended. Transfer studies are complex and sometimes even impossible because the real-world situations do not permit the objective measurement of performance of former learners. And even when these real world measures can be collected, it remains questionable to what respect the training has contributed to that performance level, and to what respect performance and performance differences can be attributed to other factors. However, it is possible to get a reasonable insight in the Transfer of Gaming, or training value of games, by means of smart experimental designs. (Korteling, Helsdingen, Sluimer, van Emmerik, & Kappé, 2011, p. 20)

Some studies have valid proof in a limited synthetic environment such as playing a modified version of sim city (Loh, Sheng, & Ifenthaler, 2015 p. 345). Others have proven that mobile technology displayed similar effectiveness as traditional learning techniques (Hwang & Chang, 2011). There are numerous experiments showcasing limited information

C. SERIOUS GAMES

Serious games have been a proven learning tool for decades (Rawitsch, 1971). Just like other niche technologies, the public and end-user do not necessarily understand its description, the capabilities, limitations, or even the proper applications. LVC dealt with similar issues of fragmented definitions in recent years. Just as LVC meant different training possibilities to different stakeholders depending on service and community, serious games represent conflicting definitions or capabilities which exacerbate the challenge to intelligently inform prospective military customers of their value. Serious games need not become the next buzz word that offers a transformation in the science of learning over traditional methods employed today. The following section defines types of games, when it’s appropriate to use them, characterizing intended goals, establishing how to adapt and personalize the SG, and defining player experience.

1. Serious Games Defined

A serious game is defined as a digital game not with the primary purpose of pure entertainment, but with the intention of serious use as in training or education (Loh et al., 2015 p. 6). This captures that there will be players using a digital interface to generate or improve comprehension of a topic.

A common error is to assume an SG is the same as gamification. Gamification leverages game mechanics to motivate the users toward certain behaviors or practices; adding game elements to non-game topics. An example is a teacher giving stickers for participating in an elementary school classroom. Serious games are designed from inception to meet specific goals to increase performance or knowledge for the user.

Characterizing the goals during the design phase of the SG is paramount. There is a direct link between SG motivation and the goals incorporated. Goals related to the design of serious games break down into six competence domains (Wiemeyer & Hardy, 2013). This thesis will focus on cognitive and perceptual competences. Subsections of cognitive and perceptual competences include planning, problem-solving, and strategic thinking.

Adaption and personalization help define SG with the expressed intention of being attractive and effective in engaging the player. Many options are available to make an SG attractive, such as having the ability to produce a unique avatar. This is also referred to as adaptability. Another important trait is adaptivity, which is monitoring the player to keep them on task via in-game assessments.

The gaming experience sums up the previous discussion but from the output side of the equation. Flow (Csikszentmihalyi, 1991) is the balance between the player's skill level and task difficulty. If the experience bores the player, or is too difficult, then the game has a poor flow. Flow must adjust for player experience, knowledge retention, and expected skill attainment. Flow need not retain a simplified one to one ratio, but the SG must incorporate enticements to bring the player toward the end-goal. An important byproduct of flow is the player's emotional state from gameplay (Novak & Johnson, 2012). Just as the designers should attempt to keep players' flow balanced, emotions must be a consideration, as depicted in Figure 2.

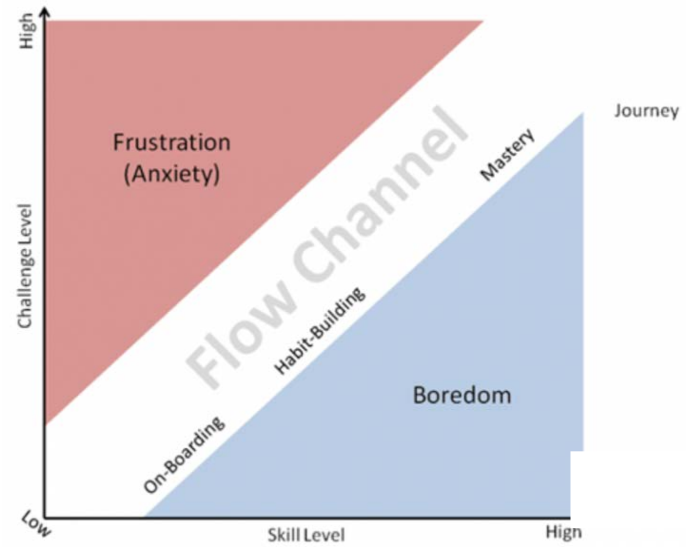


Figure 2. Flow and journey. Source: Marczewski (2012).

2. Divergence of Serious Games Design from Entertainment Games

Digital games for entertainment have been made for decades, resulting in an industry standard for rules, mechanics and gameplay design. SG designs are different than regular entertainment games, just like their intention and purpose are different. Serious games are most successful when created from scratch to engage the trainee in the desired ways. If designers transform an already built game into something outside the original design and intent, it tends to be more work without the desired effects. In comparison to entertainment games design process, an SG must account for data collection, assessment, and the user experience.

Players in entertainment and serious games need properly communicated rules to understand to build trust and operate in the anticipated design of the game. Rules provide virtual constraints and limitations so players can work to win within the boundaries. One example is the physics-modeled speed of a bullet versus the speed of an agent. They also promise greater satisfaction from the player perspective upon finishing the game (Tekinbaş & Zimmerman, 2004). When a player wins within the confines of the game and its rules, they feel a sense of accomplishment; serious games can add learning the designed material on top the entertainment value.

Mechanics and aesthetics are key aspects of any digital game, including serious games. These give a game its unique style and feel. There are many command and control (C2) serious games, including Command, Modern Air/Naval Operations. By making the player feel like they are in charge with many options to accomplish a challenging mission showcases a quality design. Graphics in serious games may emphasize an essential learning point by increasing the periodicity of specific markers on a virtual path the student needs to pick up, and reduce the number of markers later after more practice.

The user's experience (UX) refers to how seamless or fitting a game is to accomplish the required goals (Lazzaro, 2004). A SG wraps itself in many scientific fields to be successful. Figure 3 displays how the fields of science impact an SG. Programmers are crucial, but are not the only profession necessary. The UX cannot be created without key contributions from other unique fields like psychologists, instructional designers, and specialists in the gaming field. They all contribute to building an enduring, successful SG.

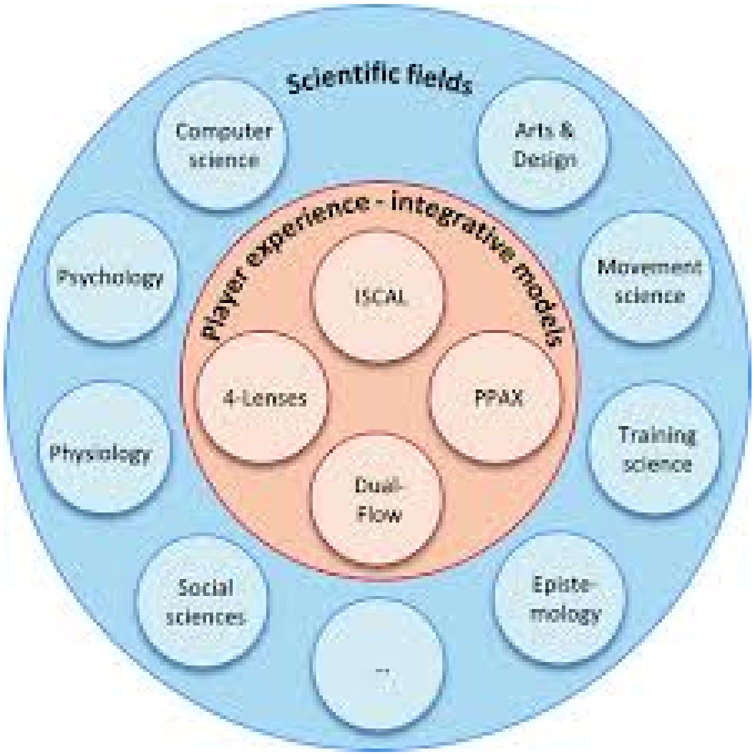


Figure 3. Integrative models of player experience. Source: Dörner, Göbel, Effelsberg, and Wiemeyer (2016).

Entertainment and serious game data collection are accomplished several different ways, and it is up to the design teams to determine how they are going to proceed. Some of the following methods may also be employed by games for entertainment, but are meant for serious games. To obtain the desired data, an SG design must identify what behaviors and performance(s) will reveal a change in the subjects' knowledge and skills. The right tasks or situations must be built to produce those behaviors. Designers refer to data collected from data logs within the game as in-situ data collection. External collection methods are known as ex-situ data collection. Ex-situ refers to the SG player's teacher or administrator injecting questions during or after gameplay. This was the typical method to collect data, but is more susceptible to bias. In-situ collection is preferred, but costlier and more time consuming to integrate with the SG programming. Without data collection, there is no way to know if any research will produce a statistically significant result.

The primary in-situ collection methods are via telemetry, evidence-centered design (ECD), and stealth assessment. Telemetry obtains and utilizes measurement data. Game telemetry is the data associated with specific game events, the state of a game, or other parameters of interest (Dörner et al., 2016). The term implies the capture and logging of game events that occur. ECD (Mislevy, Geneva, Riconscente, Rutstein, & Ziker, 2017, pp. 19–24) is an entire framework that captures and analyzes behavior log data by incorporating cognitive science, statistical modeling, and the latest SG design technologies. It works in three stages, the first is domain analysis, then domain modeling, and last is conceptual assessment framework (CAF). Finally, stealth assessment suggests an invisible, assessment tool that interlaces directly into the gaming environment. It captures real-time data, and has provisions for adapting learning based on that data. All of these in-situ methods still require an experienced person to sift through all the game data to find the meaningful results. The output can be used to improve the user experience, the design, refine teaching goals, and many other uses.

3. The Player's Perspective of the Design Process

The player's perspective includes discussion on player experience (PE) and also different player types. PE is different from UX. PE relates to the behavioral, social, and psychological level of the individual. The expectation is for the test subject to have an experience, or interaction with the SG, not just be a passive bystander. PE normally divides into three specific facets when playing: challenge, tension, and immersion. Challenge is about whether the game engages the player to try their best. Tension deals with the players ability to finish the current game objective. Immersion is a mix of enhancing realism and consequential interactions in the gaming environment. Designers of serious games must accommodate PE through narrative paradox, PE measurement, and impacts of multi-PE.

A story or novel attempting to reconcile with a game is called narrative paradox. Stories are static while games like chess rarely ever repeat in the exact same steps. Interactive storytelling is an attempt to provide the player freedom while still pursuing goals. Narrative paradox has three dimensions—simulation, ludology, and narratology. Simulation is the game type or world. Narratives help relate player’s freedom versus the designer’s intention. Ludology is the study of gaming related to player action and designed events. Narratology is the study of structure, function of themes, and associated symbols. High simulation is an avatar world. High ludology example is Tetris while high narratology is a movie. If the balance is off, both the PE and designer suffer and the SG will not be used.

PE measurement involves experimental techniques involving behavioral, physiological, and subjective methods (Dörner et al., 2016). Behavior methods can be assessed through game logs and reaction times. Physiological models comprise items like heart rate and muscle activity with supplemental technology. Subjective models include questionnaires and interviews meant to assess the player’s perception after using the SG. Questionnaires obtain information on topics from player curiosity to spatial presence to overall game-experience. Some PE measurement can be done with game metrics, persona modeling, or eye tracking. Measurement, like data collection, is crucial to obtaining usable information to answer research.

Intelligent design helps associate different types of players to their actions. This taxonomy (Bartle, 1996) is considered applicable to both single player games and multiplayer, including offline and online games. The four types are killers, achievers, socializers, and explorers. Killers prefer competitive games, pitting their skills against others. Achievers enjoy completing the entire game and any bonus material. Socializers, as the name implies, make it their primary objective to gather online friends and relish in a wide social network. Finally, explorers seek out any hidden sections or places in a game with a map. They move at their own pace, and fair worse in timed sections of games. Without knowing and constraining the audience, the SG will lose impact and it learning content will not meet the objectives.

Multiplayer serious games present additional design challenges to an already complicated single player SG. Multiplayer SG breaks down into three basic category types: competitive, cooperative, and collaborative. Competitive play means everyone fights for themselves only for the entire session. Cooperative play is a team concept for the session while collaborative play intertwines the players individual needs with temporary team-oriented goals. When mixing different learning preferences, personalities and types, the designer is attempting to bridge many combinations that could quickly nullify any possible positive outcome. The designers should precisely detail the interrelationships and interdependencies or the goals will unlikely be met.

D. PREVIOUS SERIOUS GAMES RESEARCH FOR THE U.S. NAVY

The Navy has explored the use of serious games just a few years ago. A Naval Postgraduate School (NPS) wargaming team of students in support of NWDC assessed operational level of war (OLW) though an analysis of serious games available in 2011 According to Jeffrey Appleget (email to author, June 8, 2018), they evaluated if any training tools across the DoD perform the requirements necessary to increase proficiency of mid-grade deployable staff officers. They reviewed seven serious games; some were commercial, and others were produced for government use. Three related issues to this thesis were the valuation system used, the scoring process and metrics, and measures of effectiveness (MOE).

The first input for stakeholders was to have a system to measure valuation. Leadership dictated a need which set a precedent indicating there was a training gap in the staff officer curriculum. After soliciting feedback from staff officers, looking at METs, and comparing with similar Army and USMC training systems, the OLW team was unable to achieve consensus regarding a set of training metrics common to all maritime staff officers. They decided to use NWC’s maritime staff operators’ course (MSOC), where the content attempts to capture the mindset of the mid-grade officers’ for OLW.

Once they achieved consensus, the metrics compared the serious games against each other. The six categories for evaluation were doctrine, operational planning, decision making, feedback, utilization and flexibility. Each category breaks down further to trace each subcategory to a referenced publication or instruction, assuring unbiased criteria. Some of the subcategories were ambiguous, putting the onus on the player to fully comprehend and grade it properly.

MOE were the output of assembling the survey feedback, and displaying descriptive statistics and figures. Figure 4 is an example of a descriptive output in graphical form.

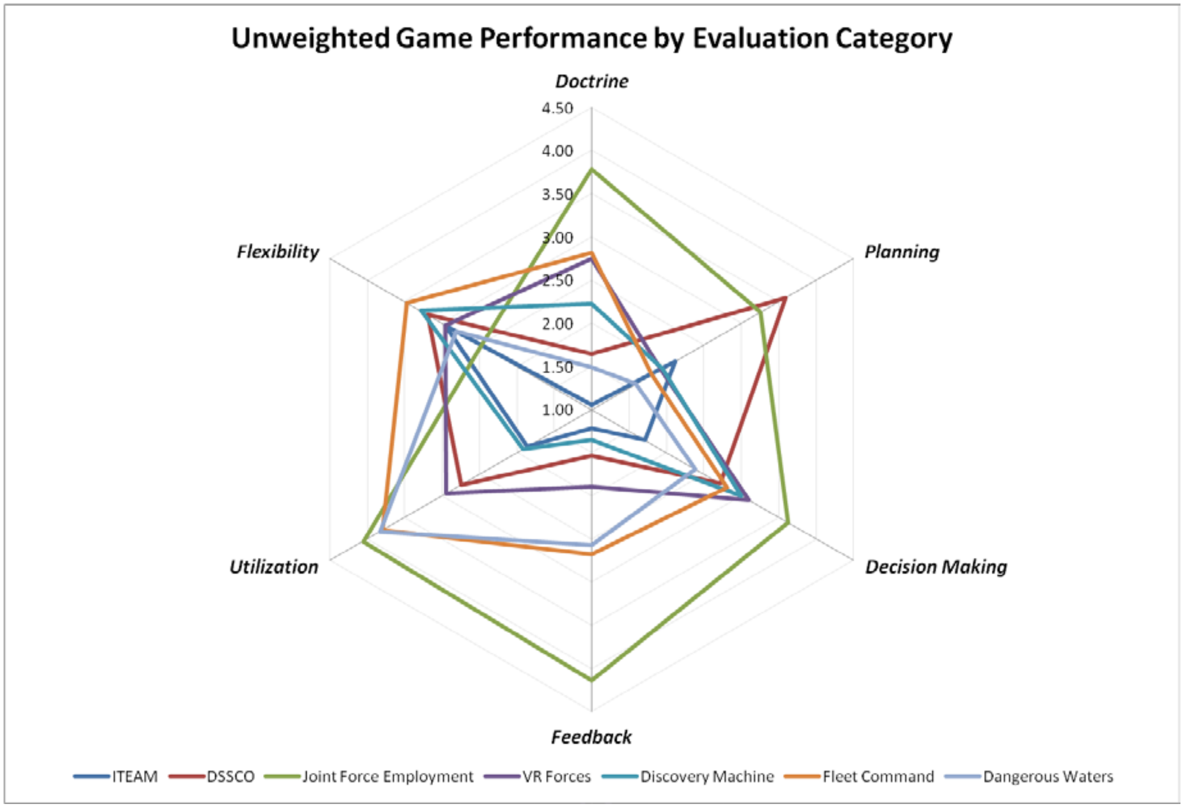


Figure 4. Example of radar graph. Source: McDowell (2016).

This radar graph provides a comparison of the SG based on the metrics outlined. This is just one method to display data collected, but shows average scores across the entire Likert survey employed. The reader can quickly interpret overall strengths and weaknesses of each game. Adding a weight to each subcategory can prioritize certain items that are of greater value to the staff officer.

E. USE OF SERIOUS GAMES FOR DEPLOYABLE STAFF TRAINING

A serious game provides the user a productive, educational interaction that should also be entertaining and motivating. As technology evolves with the science of learning, the U.S. military has shown greater interest in serious games. An SG created from scratch maximizes its benefits. However, it is common practice to leverage current technologies in order to save time and money from having to enter to procurement process from the beginning. This experiment attempts to find the closest products answering similar needs with the ability to tweak the coding to meet the training needs.

The Navy staff officer has a burdensome list of daily, weekly, and OFRP requirements to accomplish. Fitting in yet another requirement is not the goal of this thesis. However, it is reasonable to believe that augmenting some of the time devoted to reviewing instructions and publications they are responsible for could be of equal benefit or even a better use of time. The two serious games used in the thesis are jCORE and LITMUS Warfighter plug-in. They both served the Navy in different capacities as an SG or as an analysis tool.

LITMUS and jCORE possess capabilities unique to the Navy and a strike group staff. Both games previously operated more at the tactical level, but adjustments trended the focal point toward the operational level of war (OLW). There was greater emphasis placed on exercising knowledge of doctrine, decision making, and pressing to better appreciate the nuances of interaction between composite warfare commander staffs. Both serous games still generally play at the tactical level, meaning the player chooses how to employ individual units or aircraft with the intent to think as a staff officer would think. Some planning aspects are incorporated, but it is not meant to be a dedicated planning tool. There are other tools already developed for that like Athena.

Staff officers have many new skills and knowledge requirements and less time to learn them versus earlier in their career when learning to drive ships or submarines, or fly aircraft. Learning how to perform up to standard in the new billet requires having the training tools, information sources, and time to absorb and practice those skills. Most staffs, like ships or squadrons, build in time to train and review materials to improve or maintain performance. Leaders naturally want their team to operate at as proficiently as possible. Since there is not dedicated part task trainer for staff officers, the newest watchstanders are beholden to FST opportunities or underway time to practice their skills. This experiment attempts to determine the value of an SG, augmenting the training curriculum similar to an operator using a PTT or sim.

An academic point of contention related to training and readiness is proficiency. Proficiency is generally accepted as an advancement in knowledge or skill. This implies more than just an exposure to a topic, or practicing a complicated process nine months ago. Measuring proficiency is naturally even more challenging since there is not specific definition, process to follow, or standardization for how often to practice. An SG can be a tool that offers consistent availability and potentially a tailored training regimen. Progress in an SG can be measured (Scheldrup, 2018), as was the case using jCORE.

Given the previously discussed ability of serious games to address specific training needs, we decided to investigate whether serious games could have an effect upon the training. The goal of our experiment was to determine whether a serious game can improve staff members' performance on tests and get feedback on what factors of the serious games were required to improve performance.

jCORE is a browser-based gaming tool built for PMR-51 by the game design company Pipeworks. An earlier component of jCORE was named Strike Group Defender (McDowell, 2016). This SG won top prize at I/ITSEC in 2014. jCORE operates as a single or multiplayer game from red or blue force perspective. Embedded are tutorials increasing in difficulty and complexity with performance metrics including scoring by how well the player properly defends their assets from enemy missiles. Recent upgrades challenge the player to work through multi-axis problems with land, sea, and air assets available, and employing offensive and defensive measures. jCORE used unity game engine as the player interface. In total, it took about 1.5 GB of memory on the laptops.

The other SG is based on the LITMUS simulation, developed by Naval Warfare Development Center (NWDC) Dahlgren. LITMUS Warfighter plug-in leverages the analytic simulation to offer a browser-based multi-player SG. There is a tutorial to install and learn the basics of the SG so a player can minimize time learning how to play and focus learning or reviewing. Though still in a beta phase, LITMUS Warfighter plug-in offers red versus blue forces anywhere in the world to practice surface and air TTP, theatre geometry, and decision-making. LITMUS used Unity game engine as the player interface. LITMUS took up about 1.5 GB of memory on the laptops as well.

THIS PAGE INTENTIONALLY LEFT BLANK

III. APPROACH

A. HYPOTHESIS

Given the goal of determining whether SGs improved subjects' knowledge, we formulated the following hypothesis:

H_O: There is no difference in pretest and post-test one from the experimental group

H_A: There is a statistically significant increase in score from pretest to post-test one regarding the experimental group.

We later evaluated similar hypotheses by comparing the experiment subjects' scores on the pretest to those on post-test two, and those on post-test one to those on post-test two. The same format was used on the control group test scores. Finally, we compared the test scores between the two groups.

B. DESIGN OF EXPERIMENT

1. Setup

The main pieces of the experiment were the hardware, specially designed scenarios within jCORE and LITMUS, the phases of execution, and data collection including surveys and tests. This thesis consists of three major phases. Phase I was a pilot study enlisting the Navy staff SMEs. Phase II was composed of participants making up the experimental group whose training consisted of playing the two serious games. Phase III was the control group, who were trained in the traditional manner. We conducted all phases in accordance with a protocol approved by the NPS institutional review board (IRB).

The hardware used for the experiment was eight laptops, a WIFI router (Phase I), two switches (phase II), and associated cables. Four laptops contained jCORE and four had LITMUS uploaded. The jCORE laptops were a mix of Alienware and Omni brands with sufficient computational power containing Core i5 processors or better. The jCORE laptops were connected together via Netgear switch and Cat V cables. This only required setting one laptop as the server and setting proper IP addresses on the others. The four computers with LITMUS were Dell laptops from the NPS simulation experiments & efficient design

(SEED) laboratory. All four were connected in similar fashion to the jCORE laptops via a Netgear switch and cables. They also had sufficient computational power (Core i5 and above) to operate LITMUS warfighter-plugin. None of the eight laptops ever glitched or failed running their respective programs due to lacking computational power.

There were two scenarios, one in the South China Sea (SCS), and the other in vicinity of the Strait of Hormuz (SOH). There was a background political-military story that led to potential conflict. A scene setter was developed to give to blue and red force players a commander’s intent, rules of engagement (ROE), order of battle (OOB), and win/lose criteria. This scenario was designed to push blue forces into conflict since red forces quickly hit the briefed trip wire. In order for blue to win, they have to aggressively develop and implement a plan to use air assets to discern the location and attack those red forces breaking through the exclusion zone. Losing the aircraft carrier in either scenario meant blue forces lost and the session was terminated immediately. Appendix C covers the details.

The SOH scenario was composed in similar format with a background political-military situation, and scene setters laying out goals to accomplish with the given assets. This scenario was designed to address how to counter small boat attack and exercise restraint. Blue forces are split on each side of the SOH to complicate the theatre geometry, communications, and decision making. Day-one was intended to have kinetic effects only after red forces initiate offensive action. Day-two is meant to test the players ability to realize that red forces displayed hostile intent but did not meet criteria for ROE. Appendix C covers the details.

2. Phase I

a. Procedures

Phase I consisted of a pilot test where SMEs validated the tests, scenarios, and surveys. It was conducted from 14–15 January 2019. Members of both the jCORE and LITMUS development teams assisted with the pilot test. The participants played each scenario in both serious games, thus playing a total of four times. We followed the procedures according to the script during the actual experiment, and realized some improvements for Phase II. Those lessons were implemented into the final version of the experiment for Phase II.

Phase I began with participants completing the permission forms and a demographic survey. The participants were broken into two groups to play LITMUS and jCORE simultaneously. Prior to playing each game, participants received a one-hour tutorial. While playing, the subjects battled each other in the scenarios: one or two subjects commanded the blue forces, while another commanded red forces (opposition). When there were multiple participants playing on the same team, we allowed them to decide how to split control of their units. After two scenarios using the same game on the first day, the next day they conducted the same two scenarios using the other game. The scenarios were expected to take 1.5 hours, but only took one. Both game designers incorporated our feedback and other SMEs to improve the training value to the target audience.

b. Participants

Participants in Phase I were staff members of the assessment staff on the west coast. It evaluates and provides training to the staffs of carrier strike groups, amphibious groups, and expeditionary strike groups homeported in the Pacific theater. They are the acknowledged SMEs, both for strike group staff operations and training.

Five players provided inputs as SMEs. Four were lieutenant commanders (LCDRs)/O-4 and one fire control chief (FCC)/E-7. Each individual had completed many deployments before reporting to the assessment staff, and most had evaluated other staffs undergoing work-ups as a member of the assessment staff.

They were randomly divided up to play jCORE and LITMUS. A low sample of volunteers due to external operational commitments prevented being able to play in a hierarchical fashion – some players report to a player acting as the overall officer in charge. The pilot study had players on both sides instead of playing together as the blue forces. The researchers collected data on their comfort levels with video games, of which few had much experience or comfort. Those that scored their overall comfort higher with video games were quicker to understand the benefits of hot keys offered in jCORE.

3. Phase II

a. Procedures

The experiment was conducted on 27–28 March 2019. The methodology was similar to the pilot study with a few notable changes. All participants played as blue forces and the research associate played as red forces. The pretest was given prior to the tutorial session and post-test one was administered upon completion of the second session on day-two. Post-test two was sent approximately one month later. An updated survey was given to glean additional demographic information. Many of the SG suggestions were incorporated pro bono, making the gaming experience better for the players.

b. Participants

Phase II worked with a west-coast CCSG, a deployable strike group staff, at a TTGP building in Point Loma THIRD FLEET complex. Six participants spanned ranks from chief warrant officer (CWO) 2 to commander (CDR). The CCSG committed six participants to enable manning of three roles in the experiment for blue forces: the admiral (Bravo), strike commander (Papa), and Sea Combat Commander (Zulu). All participants had a wide variety of gaming experience and years of Navy service. They also filled out surveys after each of the four gaming sessions like Phase I participants, but here they completed the three tests from which Chapter IV derives its data.

4. Phase III

a. Procedures

The control-group consisted of many different staff officers using the current standard of traditional learning. TTGL delivered one week of traditional training, and allowed participants enrolled in JMTC to participate. JMTC academics were conducted in March 2019. All volunteers completed the pretest before academics commenced, and post-test one upon completion of the training. Approximately a month later, post-test two was given. No surveys were provided since nobody played an SG. However, some basic military demographic data was collected.

b. Participants

Phase III was the control group of 18 test subjects. These volunteers were attending JMTC at TTGL and volunteered to help out with the experiment as the control group. Test subjects spanned in rank from CWO2 to CAPT and years of service from 8 to 23 years. They did not take surveys, or provide gaming experience. Most of the volunteers completed all three tests.

5. Data Collection

We conducted surveys after each gaming session to record their opinions and experience. The survey's graded metrics were doctrine, operational planning, decision making, organizational construct, and feedback. Those metrics are broken down into subcategories and were all equally weighted. Players scored them from one to five. Four surveys from each participant were taken from both Phases I and II.

The three tests were used to answer the whether the hypothesis was accepted or rejected. Each test was composed of 30 questions and each volunteer was given a maximum of 30 minutes to complete. The tests were interchangeable to ensure equality and could be completed in any order. Half the questions came from TTGL and the other half I created with assistance from SMEs. Questions were multiple choice, fill in the blank, and true/false. Though some of the questions were similar, all 90 questions were unique. One question from the pretest and post-test 1 were thrown out. In one instance, the true/false question was too vague and could be interpreted so each answer was correct. The other removed question was a fill-in-the-blank. It was too complex and nobody got it entirely correct. Post-test 2 removed the question that received the most wrong answers to stay consistent with the other tests. The experimentation and control group received the same three tests, but in different order to avoid confounding the data.

C. LIMITATIONS

1. Gaming Software

Both serious games were not commercial products and thus had not undergone the traditional processes to produce to a wider audience. They both have alternate uses and for this experiment were considered beta versions. During Phase I, LITMUS crashed and had to be restarted multiple times. However, the surveys did not suggest a pronounced negative impact relative to jCORE. Phase II incorporated much of the suggestions from Phase I, but there was not time or budget to transition either SG from a more tactical gaming feel to the OLW desired.

2. Participants

Phase I had one individual have to leave after the first day, leaving only four for the second day. Such a small ‘n’ (sample size) did not compromise the pilot phase since no test scores were taken.

Phase II also had a small n due to a number of unexpected events pulling potential players away. Though nothing could be done to increase the numbers, the volunteers present were engaged with the experiment for the two days. The numbers did not decline enough to negatively impact the study.

It is expected that some participants would not complete post-test 2 since they become unavailable one month later. As stated in the IRB, everything about the experiment is voluntary.

3. Research Associates

Research associates were used for Phase II. They did not know the games prior to the pilot phase or experiment. Hours of game testing helped improve their proficiency enough to play the red forces. Due to lower than expected number of participants for Phase II, one associate played on blue forces team.

IV. RESULTS

There were two types of data collected—subjective data from surveys, and objective data from tests. The survey data was inputted into Microsoft Excel. The test data used the open source programming language for computing statistics called R studio was used. Our statistical analysis considers smaller sample sizes to compensate for not meeting the central limit theorem tenets.

A. SUBJECTIVE DATA

The Likert survey reflects what the users felt from each gaming session. A total of twenty surveys were taken between Phases I and II. Generic demographic data included SG played, gaming experience, confidence in current billet, rank, years of service, or an overall score.

The volunteers were asked four questions, as seen in Appendix D. Figure 5 displays the player output. Decimals are there if the participant filled out the demographics section of the survey more than once and changed an answer. The noticeable difference in score was in comfort with gaming. Despite the full spectrum of scoring from one to five, most players noted that it was entertaining.

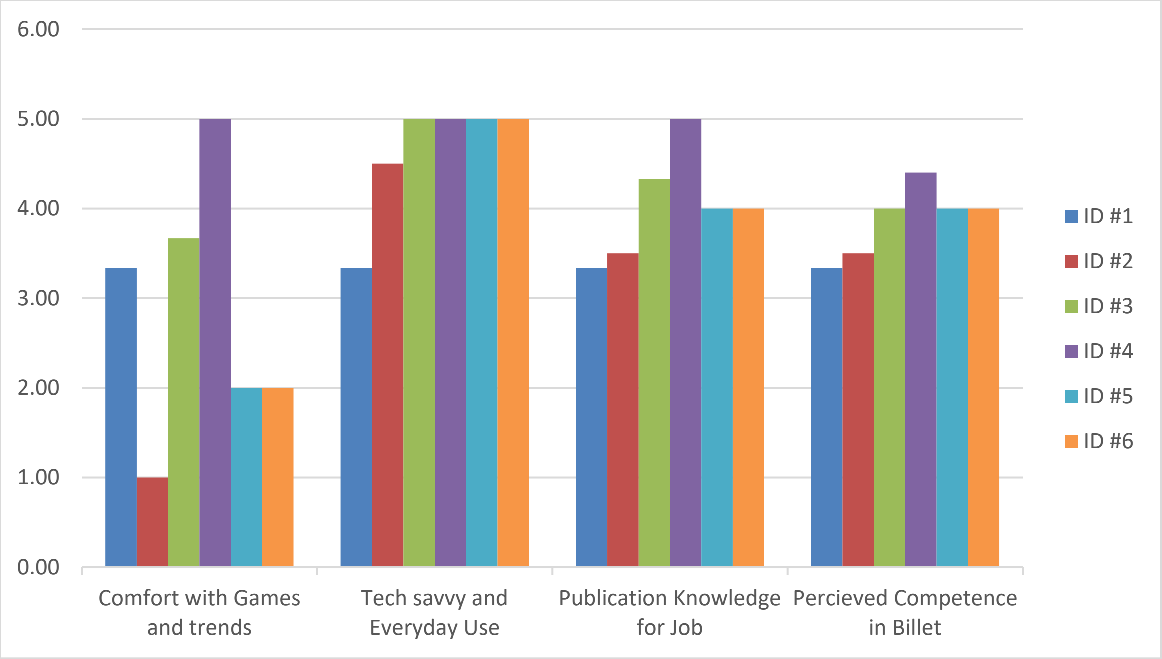


Figure 5. Experiment group demographics answers to survey

Figure 6 shows the participants scores for the first scenario played, and Figure 7 is an overall scorecard. Neither of the serious games graded out above a four out of a possible five in any of the categories. Both games were similar overall, and not as high as we initially expected since neither game was a specifically designed SG for staff watchstanders. Feedback consistently graded out much higher than the other metrics, due in part to the players better understanding the survey feedback questions.

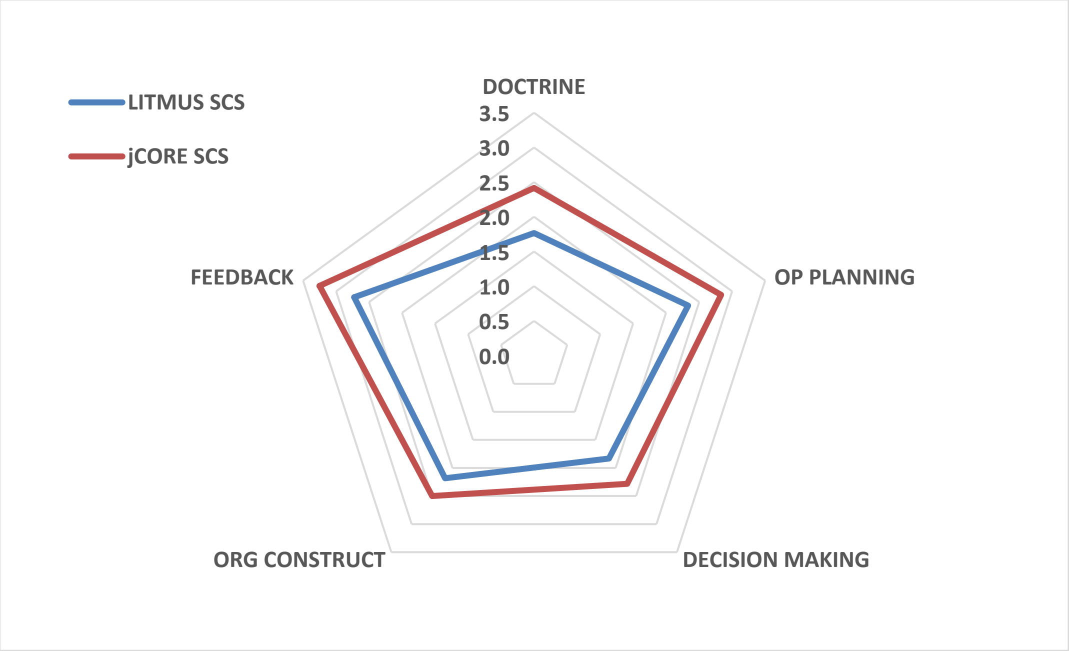


Figure 6. Phase II survey displaying results from the SCS scenario

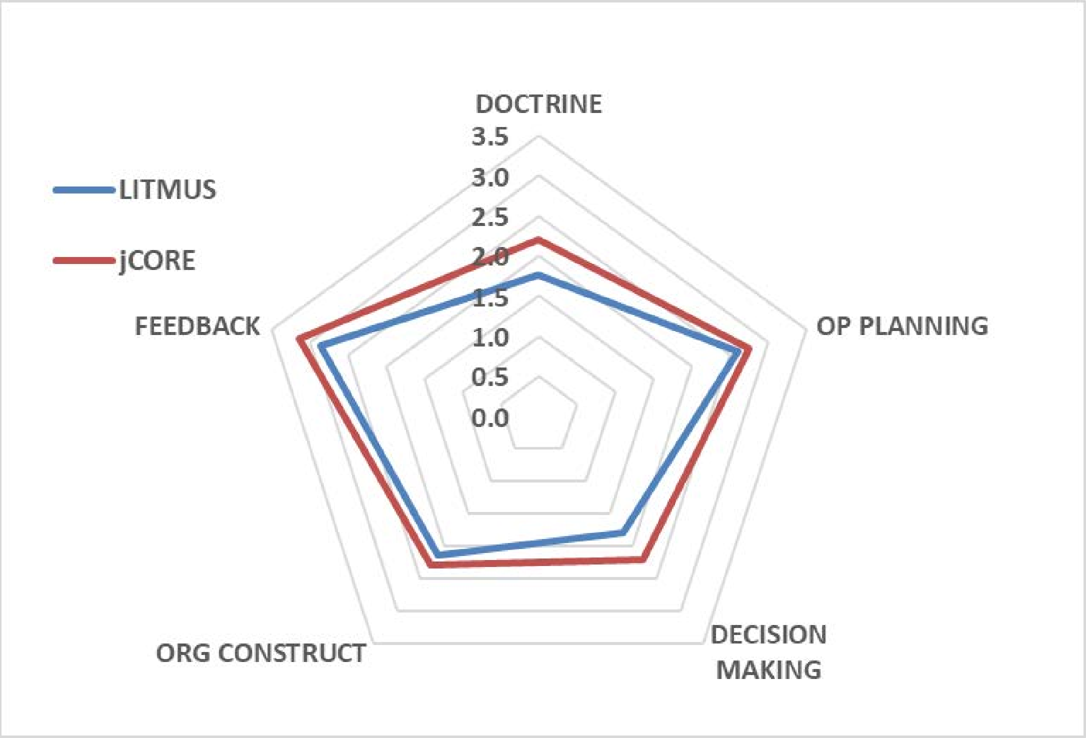


Figure 7. Phase II survey comparing overall serious games played

B. ANALYSIS OF EXPERIMENTAL DATA

‘Rstudio’ imported data from Excel as a CSV file. Initial analysis of the data sets was running the repeated measure analysis of variance (ANOVA) test. The next runs were the Cox-Stuart trend analysis (RDocumentation, n.d.-a), and finished with the sign test (RDocumentation, n.d.-c). Table 1 is a box and whisker chart visually displaying the test score results from the experiment group. The pretest, post-test 1, and post-test 2 are out of 29 possible correct. ID #2 and #6 did not complete post-test 2, and scores were included as the mean of the other four.

Table 1. Phase II test results

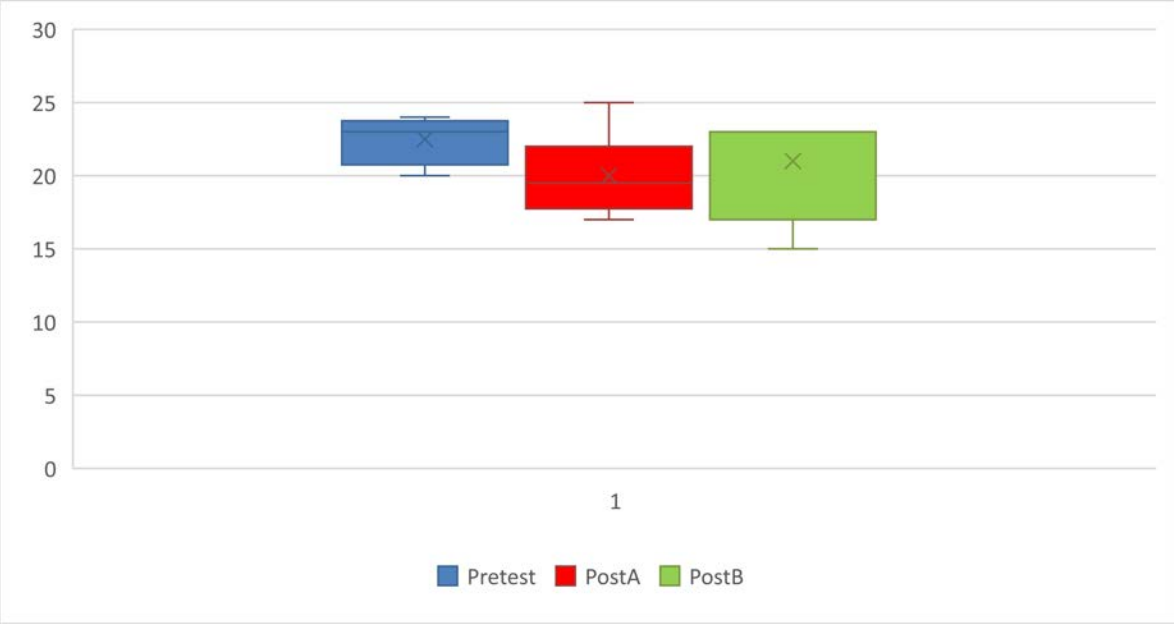


Table 2 is the basic descriptive statistics. Minimum and maximum represents the lowest and highest scored test, respectively. Median is the middle score and mean is the sum of all scores divided by the number of participants. The first and third quartiles are middle number between the minimum and maximum, respectively. The standard deviation shows variation from the means. With a small sample size, the post-test 1 minimum brings down the mean and increases the standard deviation to over double the pretest standard deviation.

Table 2. Phase II summary of descriptive data

Pretest		Post-test 1		Post-test 2	
Minimum:	20.0	Min.:	17.0	Min.	19.0
1st Quartile:	22.5	1st Qu.:	18.25	1st Qu.:	21.12
Median:	23	Median:	20.0	Median:	21.5
Mean:	22.5	Mean:	22.0	Mean:	21.5
3rd Quartile. :	23.0	3rd Qu.:	20.75	3rd Qu:	21.88
Maximum:	24.0	Maximum. :	25.0	Max.:	24.0
Standard Dev.:	1.378	S.D.:	2.828	S.D.:	1.612

There are different types of ANOVA tests. The main point for using the ANOVA test is to compare differences between means by comparing variation between the groups relative to a variation within each group. ANOVA is a method to test differences between multiple means. Inferences are made about the means through analyzing the variances. To conduct the ANOVA test, the assumptions are that the data set is a random sample, the measurements for the response in the data is distributed according to a normal distribution, each observation in the sample are independent, and the measurement is quantitative data. A large difference between the means of each group when compared to the variation within the group suggests a rejection of the null hypothesis.

The first test run was the repeated-measures ANOVA. Since the tests were not independent and there is only one factor, the repeated-measures ANOVA (also called within-subjects ANOVA) was deemed the best option. Subjects had the same comparison with different conditions since this experiment had three different tests. Another reason repeated-measures ANOVA was chosen was because it tends to have more power than the more commonly employed ANOVA version. For the repeated-measures ANOVA, and all other Rstudio coding, see Appendix A.

The null hypothesis is listed below, while the alternative hypothesis is any condition where one of the equalities does not hold up. H_{O} is the null hypothesis, E_{PRE} represents the experimental pretest, E_{POST1} post-test 1, and E_{POST2} is the post-test 2. H_{A} represents the alternative hypothesis. The significance level used was for all tests was set at $\alpha = 0.05$, and focus on “right sided.” The rationale was that we were looking to see whether SG post-test 1 would improve upon the pretest.

H_0: $E_{Pre} = E_{post1} = E_{Post2}$

H_A: one of the equalities is different from H_0

Df is the degrees of freedom. The key metric from Table 3 is the Pr(>F), which is the ‘p-value’, or probability of rejecting the null hypothesis. We related 0.18 to 0.05 and can see there is no significant difference statistically between the experimental group’s three test scores.

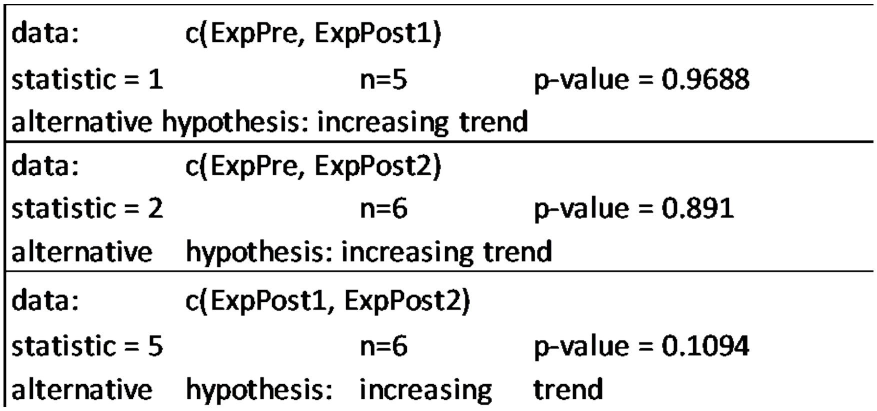
Table 3. Experimental group repeated measure ANOVA output

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
D	2	18.78	9.389	2.11	0.172
Residuals	10	44.56	4.456		

The next test run was the Cox-Stuart test. It compared each test against the other check for trends: increasing, decreasing or no trend observed. This test is valuable since it does not rely on independent data. Figure 8 shows the output from Rstudio for the test.

H_O: There is no trend.

H_A: There is a trend.



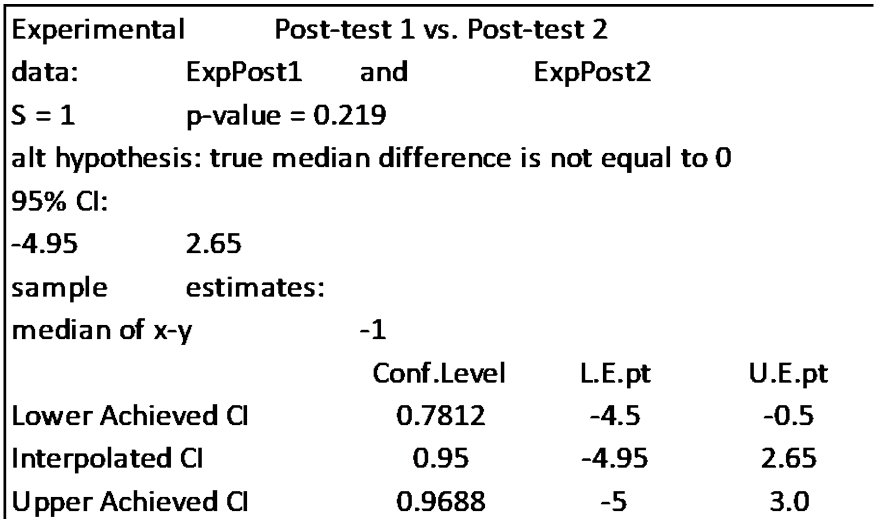
```
data:          c(ExpPre, ExpPost1)
statistic = 1                 n=5          p-value = 0.9688
alternative hypothesis: increasing trend
data:          c(ExpPre, ExpPost2)
statistic = 2                 n=6          p-value = 0.891
alternative    hypothesis: increasing trend
data:          c(ExpPost1, ExpPost2)
statistic = 5                 n=6          p-value = 0.1094
alternative    hypothesis:    increasing   trend
```

Figure 8. Cox-Stuart output comparing all three experimental group tests

The last line provides the important details of the output. An increasing trend means that results from the former to the latter increased. The p-values are all above 0.05, meaning Ho is accepted. Therefore, we cannot make any indication concerning retention of knowledge.

The final statistical test that was run is called the sign test for a two-sample paired data set. This checks for symmetry between the test scores. The sign test is non-parametric, so it does not have to fall into a particular distribution. The null hypothesis is the difference between medians equal to zero, and the alternative hypothesis is that it is not.

The key take-aways for the sign test can be found in Figure 9. The p-value above 0.05 signifies that the null hypothesis was accepted with the true mean difference not equal to zero, which was similar to Cox-Stuart trend analysis. The other sign test comparison outputs are in Appendix B.



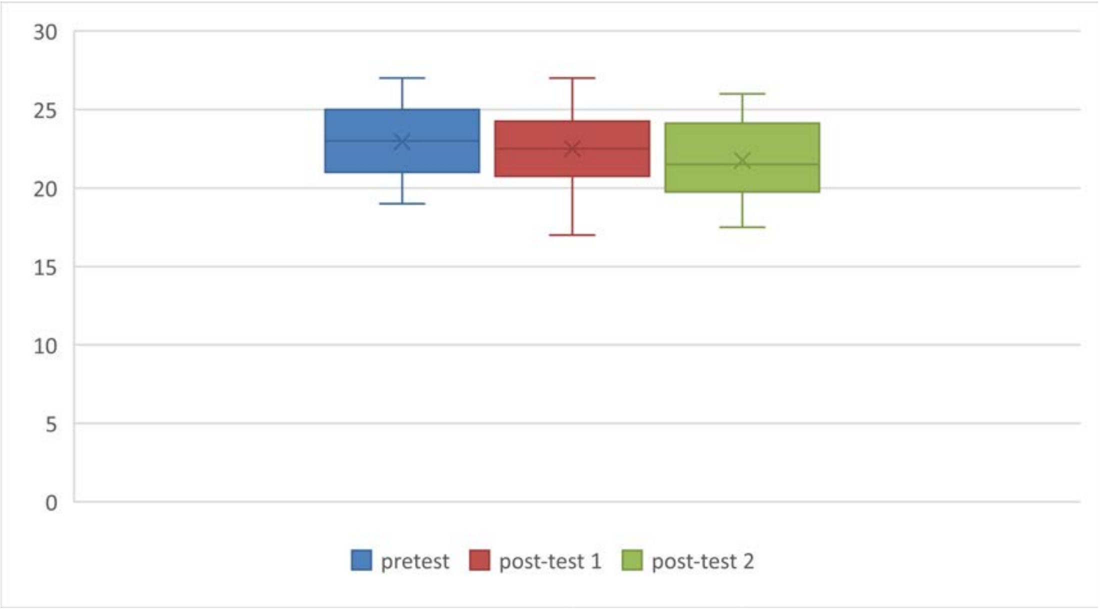
Experimental	Post-test 1 vs. Post-test 2			
data:	ExpPost1	and	ExpPost2	
S = 1	p-value = 0.219			
alt hypothesis: true median difference is not equal to 0				
95% CI:				
-4.95	2.65			
sample	estimates:			
median of x-y		-1		
		Conf.Level	L.E.pt	U.E.pt
Lower Achieved CI		0.7812	-4.5	-0.5
Interpolated CI		0.95	-4.95	2.65
Upper Achieved CI		0.9688	-5	3.0

Figure 9. Sign test comparing the experiment group pretest to post-test 1 from Rstudio

C. ANALYSIS OF CONTROL GROUP DATA

The control group was larger than the experimental group due to the fact that it had JMTC to draw upon for volunteers. Table 4 is a box and whisker chart visually displaying the test score results from the control group. Some of them did not have time to take post-test 2 one month after finishing JMTC, where they completed the pretest and post-test 1. ID #7, #8, #10 and #11 did not complete post-test 2, and scores were included as the mean of the other fourteen.

Table 4. Phase III control group test results



The values in Table 5 display the descriptive statistics associated with the control group. The scores are surprising close across the board.

Table 5. Phase III summary of descriptive data

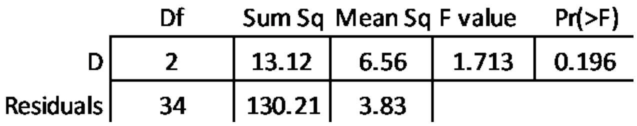
Pretest		Post-test 1		Post-test 2	
Minimum:	19.0	Min.:	17.0	Min.	17.5
1st Quartile:	21.0	1st Qu.:	21.0	1st Qu.:	20.62
Median:	23.0	Median:	22.5	Median:	21.75
Mean:	22.94	Mean:	22.5	Mean:	21.75
3rd Quartile:	25.0	3rd Qu.:	24.0	3rd Qu:	23.5
Maximum:	27.0	Max.:	27.0	Max.:	26.0
Standard Deviation:	2.338	S.D.:	2.684	S.D.:	2.325

The control group analysis follows the same statistical testing as the experimental group: repeated-measure ANOVA, Cox-Stuart test, and the sign test. The null and alternative hypothesis are similar to the experimental group, except. Here, CPRE, CPOST1 and CPOST2 represent the pretest, post-test 1, and post-test 2 respectively. The significance level used for all control group tests was set at α = 0.05. Table 6 displays the repeated ANOVA output from Rstudio.

H_0: $C_{Pre} = C_{post1} = C_{Post2}$

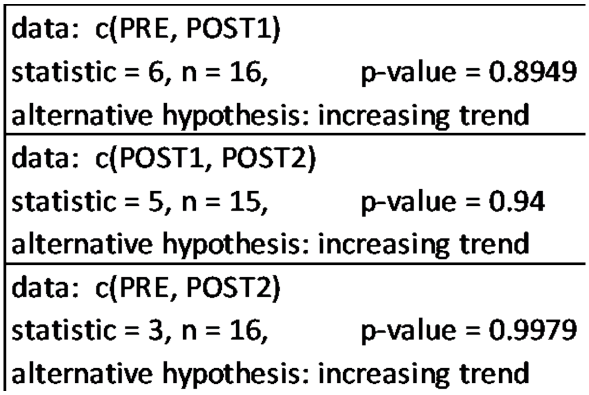
H_A: one of the equalities is different from H_0

Table 6. Control group repeated measure ANOVA output



	Df	Sum Sq	Mean Sq	F value	Pr(>F)
D	2	13.12	6.56	1.713	0.196
Residuals	34	130.21	3.83		

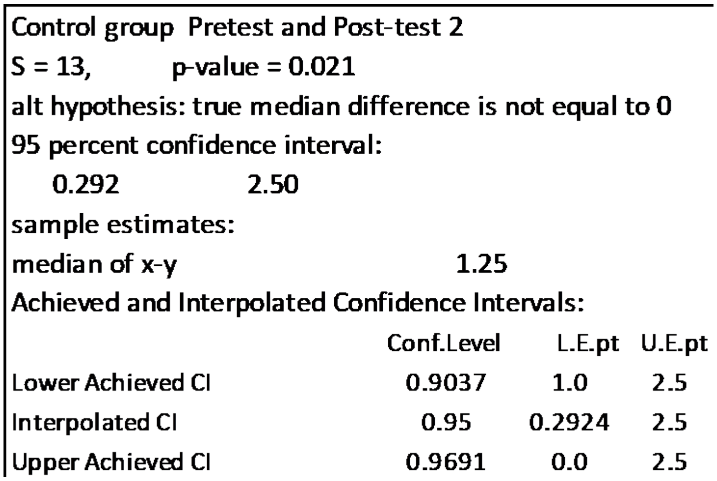
The next test run was the Cox-Stuart test, checking the control group for trends. Figure 10 shows the output from Rstudio for the test. The Cox-Stuart test 'p-value' was greater than α. None of the p-values were notable for any of the three combinations. This means we accepted the null hypothesis for the control group that there is no significant difference in test scores trending. A higher sample size (n) does not improve the p-value.



```
data:  c(PRE, POST1)
statistic = 6, n = 16,          p-value = 0.8949
alternative hypothesis: increasing trend
data:  c(POST1, POST2)
statistic = 5, n = 15,          p-value = 0.94
alternative hypothesis: increasing trend
data:  c(PRE, POST2)
statistic = 3, n = 16,          p-value = 0.9979
alternative hypothesis: increasing trend
```

Figure 10. Cox-Stuart output comparing all three control group tests

The final control group test was the sign test for a two-sample paired data set. The key take-aways for the sign test is in Figure 11. The p-value is below 0.05, signifying that the null hypothesis is rejected with the true mean difference above zero. This means we can state that the scores are significantly different after a month (post-test 2) from first taking a test (pretest). The comparison of the post-tests again revealed a high p-value similar to Cox-Stuart trend analysis. The null hypothesis was retained and accepted that there was no difference between the post tests. This cannot supply proof that there is or is not an increase in knowledge or about retention.



```
Control group  Pretest and Post-test 2
S = 13,          p-value = 0.021
alt hypothesis: true median difference is not equal to 0
95 percent confidence interval:
    0.292               2.50
sample estimates:
median of x-y                          1.25
Achieved and Interpolated Confidence Intervals:
                                  Conf.Level     L.E.pt  U.E.pt
Lower Achieved CI                   0.9037       1.0     2.5
Interpolated CI                     0.95        0.2924   2.5
Upper Achieved CI                   0.9691       0.0     2.5
```

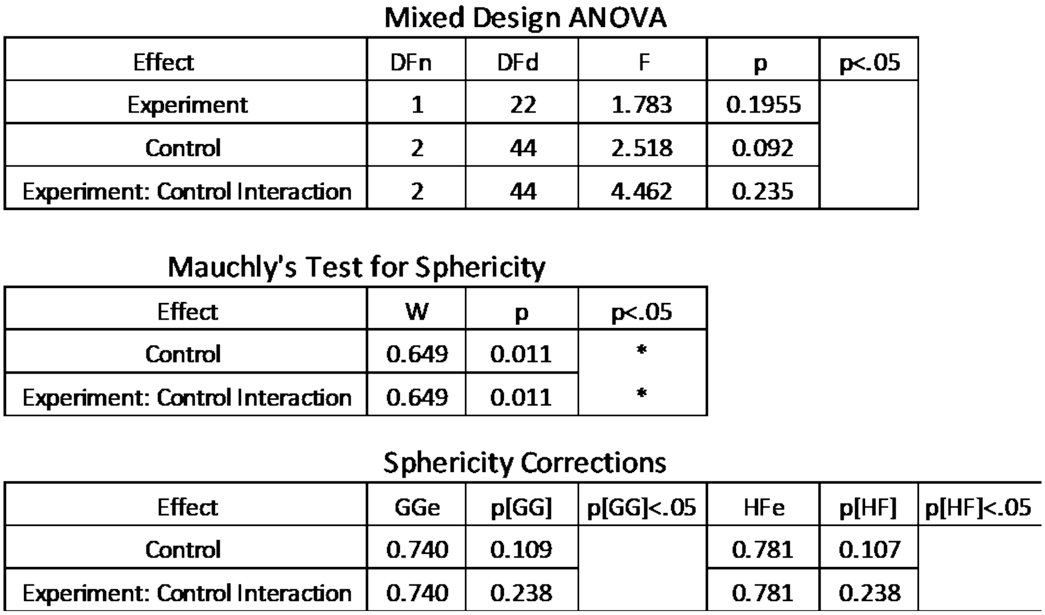
Figure 11. Sign test comparing the control group pretest to post-test 2 from Rstudio

D. COMPARISON OF EXPERIMENTAL AND CONTROL GROUP DATA

In order to compare two independent variables each with three dependent variables (tests), a different statistical test is required. The only one that is non-parametric, capable of handling smaller samples, and compare different sized groups is the mixed design ANOVA (RDocumentation, n.d.-b) Mauchly’s test for sphericity and sphericity correction test are additional, related tests checking within-subject conditions are equal.

The mixed-design ANOVA test can compare the means with repeated measures as well as independent measurements, and provide a single output, which is shown in Table 7. To understand the figure, the left column represents the groups and their interaction. DFn and DFd are degrees of freedom for the independent and dependent variable, respectively. GGe is the Greenhouse-Geiser effect, which evaluates within-subjects test output continuously. p [GG] is the p-value after the GGe is incorporated. HFe is Huynh-Feldt epsilon. p [HF] is the p-value after applying the HFe correction.

Table 7. Data comparison of experimental and control group



Mixed Design ANOVA

Effect	DFn	DFd	F	p	p<.05
Experiment	1	22	1.783	0.1955	
Control	2	44	2.518	0.092	
Experiment: Control Interaction	2	44	4.462	0.235	

Mauchly's Test for Sphericity

Effect	W	p	p<.05
Control	0.649	0.011	*
Experiment: Control Interaction	0.649	0.011	*

Sphericity Corrections

Effect	GGe	p[GG]	p[GG]<.05	HFe	p[HF]	p[HF]<.05
Control	0.740	0.109		0.781	0.107	
Experiment: Control Interaction	0.740	0.238		0.781	0.238	

Table 7 has three different p-values under the prescribed α. Starting with the mixed-design ANOVA, unfortunately the interaction’s low p-value does not provide tangible evidence about the experimental or control groups test scores. The sphericity corrections also are below α. Again, the interaction doesn’t give information that provides a meaningful proof to accept or reject the hypothesis.

V. CONCLUSION

A. DISCUSSION

Navy staff officers rely heavily on their experiences from their first sea tours. Though the same framework is used to prepare for deployment (OFRP) for staffs and ships or squadrons, the training curricula are vastly different. Staff officers get a quick exposure, some FST training, and live underway time. However, there may be opportunities to augment their busy schedule with additional training aids. The background research provided answers to the first three research questions while the experiment compared serious games to traditional classroom training from the fourth research question.

The first question asked if the training standard was met with current options. It is as the curriculum is written, however each option after the exposure course(s) has no individual review or training option; everything is a large-scale event like FST or getting a CSG underway. This points to a need for individuals to have a lower-fidelity option like a part task trainer.

Serious games must be specifically designed with the intention of meeting watchstander requirements and CSG METs in order to properly and effectively train the staff officer. Factors such as design flow, PE, UX, and data collection must be considered with purposing a SG for military training.

Serious games cannot replace the training that is being conducted, and never should alone. There will continue to be a need to train in the live and virtual environments. There does appear to be room for a part-task trainer that is easily accessible and can be completed in a short time span. Precedent was set years ago that a need exists for a lower-fidelity training system, but incorporating it into the staff officer training curriculum is more of a cultural impediment than technical.

The experiment hypothesized that classroom training and serious games would yield the different results for test improvement and retention. There were six participants in the experimental group that played two different serious games over two days compared to 18 volunteers in the control group taking JMTC for a week of classroom lecture. Three tests were administered to both groups while the experiment group also completed surveys after each gaming session. The three tests provided the objective data with the surveys enabling subjective data to be gathered. Before training began, a 30-question pretest was given followed by a post-test after training was completed. Post-test 2 was done about one month after completion of training.

The statistical analysis was conducted on the experimental group, the control group, and a comparison of independent groups. The statistics tests coded in Rstudio were the repeated measures ANOVA, the Cox-Stuart trend analysis, the Sign test, and mixed-design ANOVA test. Though the process was followed properly, only one statistically significant difference in the tests were found that showed a positive trend between the pretest and post-test 2 in the control group. All others were did not meet the threshold of p-value less than or equal to 0.05.

The surveys did yield some validating comments about serious games. SG use was positive across ease of use, engaging, and considered more enjoyable than classroom training. The last consensus was that there was a deployable staff training gap. SG were seen as a possible low-fidelity solution to train on-demand for individual watchstanders or small groups.

B. RECOMMENDATIONS

The experiment was successful in providing a platform where staff officers would like to see further investigation and experiments. Though this was a low budget, limited experiment with minimal statistically significant results, there were some worthy insights to share.

- Digital data collection. Some in-situ data collection is already present, such as jCORE’s after action review, but much more could be programmed and incorporated to coincide with pre-established goals. Knowing the metrics and end-state training objectives could tie into what data to collect from the program.

- Refined gaming experience. The players who were more proficient gamers learned much quicker about hot to defeat the enemy and did not need to use Navy doctrine or unit TTPs. Instead of open-ended red versus blue assets, a focus should be on building a lesson from a publication like NTTP 3–60.2 Maritime Dynamic Targeting.

- Single player training. If a single player version was pursued, it would need to get away from a gaming experience like Starcraft 2 when the best players click over 150 times per minute. The gaming session should be more scenario-based with three to four timed-choices, and associated consequences from those decisions with feedback.

- Multi-player training. This version would significantly contrast single player training. This would focus on soft skills like decision making, communications, and understanding the roles of each player within the strike group.

C. FUTURE WORK

The following bullets suggest where dedicated efforts could positively impact Navy training. The above recommendations could be incorporated here, but these are meant to be stand-alone spring boards into separate studies.

- Training effectiveness evaluation. A TEE consists of much more dedicated research than one student, some funding, and a couple research associates setting up some gaming sessions. This would investigate deeper into each area that was touched upon here in this experiment. A fully funded study may find more insightful, statistically significant results.

- Closed-loop simulation. Wargaming and training are worthy research topics. Another avenue worth exploring with the operations analysis department would be to employ LITMUS or another closed-loop simulation to see where bottlenecks in information and decision flow may occur on a staff or effects chain. Inserting higher or lower proficient watchstanders may prove out where to best put time to improve their skills.

- Standardization. A staff officer’s job has tasks, conditions and standards just like most any other job in the Navy. A comprehensive examination of the training curriculum and process is due. Investigation could confirm the training curriculum requires more than exposure courses, some FST events, and a few weeks of live training.

APPENDIX A. CODING FROM R-STUDIO

The following tests with the exception of mixed-design ANOVA were run multiple times. The first lines pull the test scores associated with the subjects' ID. The next lines set up each part of the necessary data to incorporate into the 'R' coding requirements. The 'R' code was pulled directly from Rstudio.

```
###########  DATA   #################
ExpPre<-c(23,23,24,20,23,22) # experimental group pretest
ExpPost1<-c(18,21,25,20,19,17) # experimental group post-test 1
ExpPost2<-c(19,21.5,22,21,24,21.5)   #experimental group post-test 2
PRE <-c(24,22,21,22,25,27,24,26,20,24,20,25,19,21,25,21,22,25) # control pretest
POST1 <- c(27,24,24,25,24,25,23,27,20,23,19,21,17,22,21,21,20,22) # control post-test 1
POST2 <- c(19,22,21,20.5,24,24.5,21.75,21.75,17.5,21.75,21.75,24,18,20,21,26,25,22)

##### DESCRIPTIVE STATS  ##############
summary(ExpPre)
summary(ExpPost1)
summary(ExpPost2)
summary(PRE)
summary(POST1)
summary(POST2)
sd(ExpPre)
sd(ExpPost1)
sd(ExpPost2)
sd(PRE)
sd(POST1)
sd(POST2)

##### repeated measure ANOVA   ##############
## exp group##
Activation <- c(D[,2],D[,3],D[,4])
Subject<-factor(rep(D[,1],3))
D<-factor(rep(c("Pretest","PostA","PostB"),rep(6,3)))
aovD<-aov(Activation~D+Error(Subject))
summary(aovD)

path<-'/Users/danielcain/TestScoresExpcsv.csv'
data<-read.csv(path)
data
D<-read.csv(path)
Activation <- c(D[,2],D[,3],D[,4])
Activation
```

```
length(Activation)
Subject<-factor(rep(D[,1],3))
Subject                          # 8-17 works
Test<-factor(rep(c("T1","T2","T3"),rep(6,3)))
aovD<-aov(Activation~Test+Error(Subject))
summary(aovD)                                   # works

            ### control group ##
path<-'/Users/danielcain/TestScoresCONTROL.csv'
data1<-read.csv(path)
data1
D.1<-read.csv(path)
Activation.1 <- c(D.1[,2],D.1[,3],D.1[,4])
Activation.1
length(Activation)
Subject<-factor(rep(D.1[,1],3))
Subject
Test<-factor(rep(c("T1","T2","T3"),rep(18,3)))
aovD<-aov(Activation.1~Test+Error(Subject))
summary(aovD)

#######  COX STUART ######

library(randtests)
cox.stuart.test(c(ExpPre,ExpPost1), "right.sided")
cox.stuart.test(c(ExpPre,ExpPost2), "right.sided")
cox.stuart.test(c(ExpPost1, ExpPost2), "right.sided")
cox.stuart.test(c(PRE,POST1), "right.sided")
cox.stuart.test(c(POST1, POST2), "right.sided")
cox.stuart.test(c(PRE,POST2), "right.sided")

##########  SIGN TEST   ##########

library(BSDA)
SIGN.test(ExpPre,ExpPost1,
          alternative = "two.sided",
          conf.level = 0.95)
SIGN.test(ExpPre,ExpPost2,
          alternative = "two.sided",
          conf.level = 0.95)
SIGN.test(ExpPost1,ExpPost2,
          alternative = "two.sided",
          conf.level = 0.95)
PRE
POST1
```

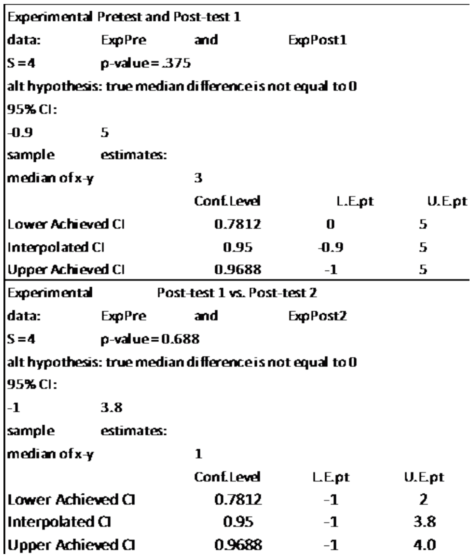
```
POST2
SIGN.test(PRE,POST1,
        alternative = "two.sided",
        conf.level = 0.95)
SIGN.test(PRE,POST2,
        alternative = "two.sided",
        conf.level = 0.95)
SIGN.test(POST1,POST2,
        alternative = "two.sided",
        conf.level = 0.95)

######  mixed design ANOVA   ########
library(ez)
TestData<-read.table("TestScoresRY2.csv", header = TRUE,sep=",")
Testdata  ## upload data set
summary(TestData)       ## Print summary
rt_anova = ezANOVA(data=TestData, dv=Data, wid = ID, within = Label, between =
Experiment)
print(rt_anova)
```

THIS PAGE INTENTIONALLY LEFT BLANK

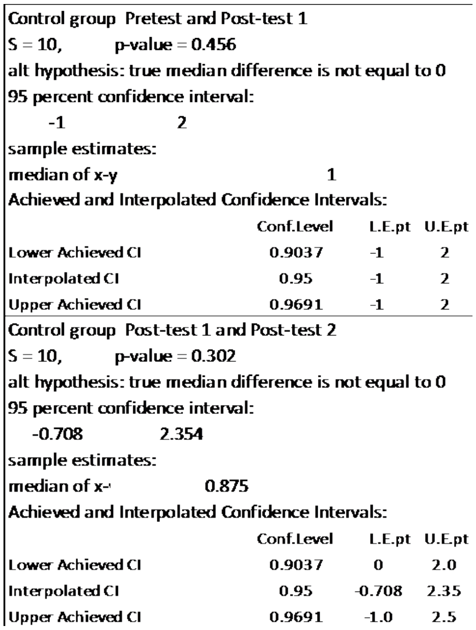
APPENDIX B. OUTPUT FROM R-STUDIO

Here are the outputs not listed in the thesis. The data here is not statistically significant but provided for context and disclosure.



```
Experimental Pretest and Post-test 1
data:          ExpPre        and           ExpPost1
S =4           p-value= .375
alt hypothesis: true median difference is not equal to 0
95% CI:
-0.9           5
sample         estimates:
median of x-y                3
                             Conf.Level         L.E.pt        U.E.pt
Lower Achieved CI              0.7812           0             5
Interpolated CI                0.95            -0.9           5
Upper Achieved CI              0.9688          -1             5
Experimental            Post-test 1 vs. Post-test 2
data:          ExpPre        and           ExpPost2
S =4           p-value=0.688
alt hypothesis: true median difference is not equal to 0
95% CI:
-1             3.8
sample         estimates:
median of x-y                1
                             Conf.Level     L.E.pt       U.E.pt
Lower Achieved CI              0.7812        -1            2
Interpolated CI                0.95          -1            3.8
Upper Achieved CI              0.9688        -1            4.0
```

Figure 12. Experimental group sign test results



```
Control group  Pretest and Post-test 1
S = 10,         p-value = 0.456
alt hypothesis: true median difference is not equal to 0
95 percent confidence interval:
      -1                  2
sample estimates:
median of x-y                              1
Achieved and Interpolated Confidence Intervals:
                                    Conf.Level    L.E.pt  U.E.pt
Lower Achieved CI                     0.9037        -1       2
Interpolated CI                       0.95          -1       2
Upper Achieved CI                     0.9691        -1       2
Control group  Post-test 1 and Post-test 2
S = 10,         p-value = 0.302
alt hypothesis: true median difference is not equal to 0
95 percent confidence interval:
    -0.708             2.354
sample estimates:
median of x-            0.875
Achieved and Interpolated Confidence Intervals:
                                    Conf.Level    L.E.pt  U.E.pt
Lower Achieved CI                     0.9037         0      2.0
Interpolated CI                       0.95        -0.708    2.35
Upper Achieved CI                     0.9691       -1.0     2.5
```

Figure 13. Control group sign test results

THIS PAGE INTENTIONALLY LEFT BLANK

APPENDIX C. SCENARIOS

A. SOUTH CHINA SEA SCENARIO

1. Political-Military Background

In 2020, RED is the world’s largest economy. However, with its newfound status, GDP has slowed to under 1% annually and recession is looming. The gap in living standards between poor, inland farmers and the more urban coastal dwellers has grown, leading to civil unrest. Many civilian analysts assume that RED’s bellicose rhetoric regarding its territorial claims is an attempt to distract the population from internal difficulties, and strengthen civil unity under the Communist Party banner.

With its burgeoning middle class, RED demands for oil and natural gas are ever increasing. Since 2017, RED has continued to fortify its claims in the Spratly Island chain, expanding land reclamation projects and now has its sights on the southern tip of the South China Sea with the island of Natuna Besar, a small island chain belonging to GREEN. Annexing this tiny island chain would give RED complete control of the busiest sea lanes in the world.

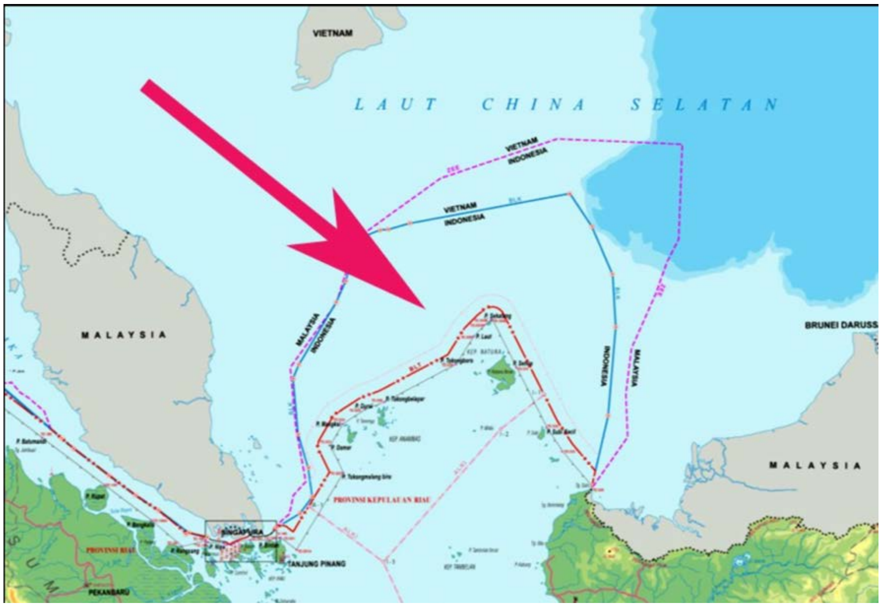


Figure 14. Location of Natuna Besar

BLUE has continued freedom of navigation operations (FONOPS) through RED's excessive maritime claims, yet this has had no effect on RED military build-up or claims. Further, RED warships have become more aggressive in their enforcement activities, and regularly harass fishing vessels of other nations, including territorial waters of Natuna Besar.

GREEN is an archipelagic nation on the southern edge of the South China Sea. Internal instability in the main islands of GREEN has simmered, preoccupying the GREEN's naval forces. It had no appreciable military ties with BLUE. Sensing a potential threat from RED, BLUE has sought to reassure regional stakeholders in the South China Sea, especially GREEN. Just in the last month, BLUE and GREEN have signed a mutual protection pact to include Natuna Besar. RED state-controlled media declared recent the recent arms sales dialogue from BLUE to GREEN as an "unfair effort" to restrain RED. RED citizens are clamoring social media, calling for a "humbling" of BLUE forces in the region. Vietnam has provided a logistics and port facilities to RED vessels begrudgingly to avoid greater conflict.

Since 2016, the BLUE Navy has adapted its fleet design to meet future challenges. Through a combination of doctrinal and materiel development, BLUE has transitioned from a "platform-centric" to a "fleet-centric" force. Therefore, air, surface, and subsurface kill chains are highly resilient, consisting of networked ships, aircraft, weapons, and unmanned systems.

2. Disputed Waters

In January 2020, GREEN fishermen, tired of harassment by RED combat vessels and no help from GREEN government, staged a series of protests. Additionally, for the last two weeks, RED warships harassed white shipping. Some of the ships were BLUE allied-flagged motor vessels in international water.

In March of 2020, a GREEN patrol craft did not return from its patrol. GREEN fisherman reported seeing a RED warship fire upon and sink it. Official RED accounts of the incident state that a DDG in the area saw the PC sinking and attempted to render aid, but nobody survived. Some in the GREEN government claim that the RED civilians have inquired about expanding Natuna airport. RED has hinted at putting ground forces on Natuna Besar in order to ensure "peaceful air and maritime operations" are abided by according to international law.

3. Political Situation

RED political will to start an actual conflict is considered imminent. RED may perceive the upcoming BLUE FONOPS and possible GREEN military acquisitions as a potential threat, giving them a real incentive to occupy before the island is fortified. BLUE is not offered safe harbor in its territorial waters, due to historically cautious relationship. Further, it is an election year for BLUE's President, and failure to show resolve during this quickly escalating foreign policy crisis may cost reelection.

4. Military Activity

BLUE satellite imagery shows a RED Amphibious Mechanized Infantry Brigade was already embarking on amphibious ships for a "previously planned" exercise. RED's South Sea Fleet has started to sortie ships and possibly submarines, and that they have increased maritime patrols over the South China Sea.

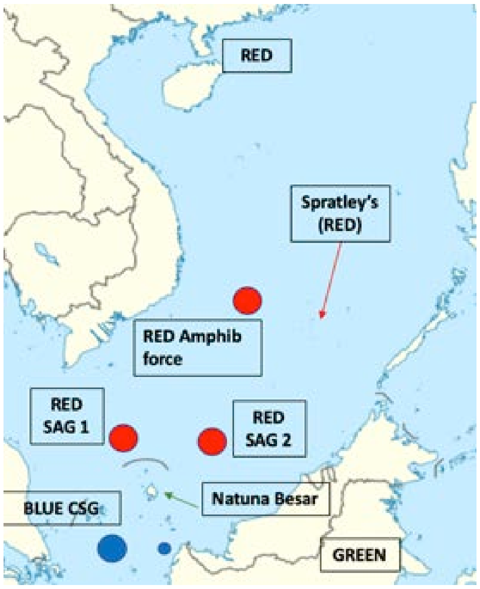


Figure 15. Intelligence summary—RED and BLUE OOB IVO Natuna Besar 24 hours old

BLUE is determined to support GREEN, deploying forces to respond to RED threats and prevent a land grab at all costs. Commander Carrier Strike Group (CSG) NINE is in the Sea of Japan conducting exercises and is scheduled to arrive within two days. The surge-ready CSG in San Diego is getting underway, but cannot arrive for another two weeks. Therefore, Commander, US Seventh Fleet (C7F) must rely on forces already in the vicinity of the South China Sea. Time is on RED's side, and BLUE must respond with forces on hand until reinforcements can arrive.

BLUE intel suggests all RED submarine threats have been accounted for via satellite imagery over the last four weeks. Additionally, a BLUE fast-attack submarine is three days out from assisting. BLUE forces have been authorized to engage should the RED amphibious force get within 50NM of Natuna Besar.

B. SOUTH CHINA SEA BLUE SCENE SETTER

1. Purpose

LITMUS and jCORE – South China Sea explore Fleet Design in naval warfare at the operational level. Like any wargame, they are designed to capture the human elements of warfare. Therefore, players will be required to make decisions in the face of uncertainty.

2. Game Design

LITMUS and jCORE are closed wargames, meaning that opposing teams will play on opposite sides of room. Each team will have knowledge of:

1. Events that led to conflict (scenario)
2. Objectives (mission goals), provided by leadership and higher headquarters
3. Own force composition and capabilities
4. Capabilities of possible enemy platforms

Each team will have incomplete knowledge of:

1. The true enemy objective
2. The enemy force composition

Teams will have to make decisions based on:

1. The intelligence, surveillance, and reconnaissance (ISR) plans they develop

2. Technical and intelligence injects

Adjudication will rest with the student researcher, based on wargaming experience, analysis and knowledge of the combat modeling tools.

Commander's intent

Prevent the landing of any one of the three RED amphibious ships on Natuna Besar.

- If fired upon, exercise self-defense with proportionality.

- Enforce the 50NM exclusion zone (EZ) surrounding Natuna Besar. Maneuver to anticipate any Red force threat without initiating effects if they have not entered the EZ.

- Establish continuous search of EZ. All assets are to bear the responsibility of, and share the duties of search.

- Any interaction outside of EZ shall be professional, and IAW UNCLOS.

- Maintain open SLOC from Japan to Australia for white shipping.

- Do not seek refuge in country Green's TTW for any reason.

- All asset Link-capable shall immediately share information up the appropriate CoC. Should there be a Link or GPS-denied environment, follow appropriate guidance.

Posture: Rd / Ti.

Mission Goals

Each force has is a set of desired outcomes from game play. Below are the details in order to achieve a "win" or "loss". Should one side not definitively achieve the requisite number of conditions, then it will each win condition is worth one point. The side with the most points wins, while a tie is a "détente".

BLUE Forces Goals

In order to win, blue must achieve three of the following while adhering to other specifics:

- Achieve a kill ratio equal of 3:1 for all assets
- Prevent RED any amphibious landing on Natuna Besar
- Kill 90% of RED air assets found
- Sink 70% of RED surface assets found

BLUE loses if any of the following conditions exist by the end of game play:

- BLUE CVN is sunk
- An entire SAG / CSG is not capable of conducting offensive operations at the end of game play
- BLUE units seek shelter from GREEN (goes inside territorial waters to avoid conflict, excluding Natuna Besar)
- BLUE player chooses wrong defensive capability vs threat over 30% of time

C. STRAIT OF HORMUZ SCENARIO

1. Political-Military Background

In 2020, RED is the largest and most powerful country in the region. Despite its status as the dominant regional leader, GDP has declined by 3% annually due to an inability to get their oil to the world market. The decline in living standards has grown, leading to civil unrest. Many civilian analysts assume that RED’s bellicose rhetoric regarding its territorial claims is an attempt to distract the population from internal difficulties, and strengthen civil unity under the elected and appointed institutions of government.

With its dwindling budgets, RED demands for oil and natural gas exports are ever increasing. Since 2017, RED has continued to fortify its regional authority to include lands west and north of the gulf, as well as threaten white shipping into and out of the straits of Hormuz (SOH). The bottleneck is fortified with strike and reconnaissance aircraft, fast attack craft (FAC), fast inshore attack craft (FIAC), land-based anti-ship cruise missiles (ASCM) and coastal integrated air defense systems (IADS).

BLUE has continued freedom of navigation operations (FONOPS) through RED's excessive maritime claims, yet this has had no effect on RED military build-up or claims. Further, RED Islamic Revolutionary Corps vessels have become more aggressive in their harassment activities of shipping vessels of other nations, including GREEN.

GREEN is a small nation on the northwestern edge of the Arabian Gulf. Though rich in oil, GREEN does not have the military capabilities to counter Red. It has strong military and diplomatic ties with BLUE. Sensing a potential threat from RED, BLUE has sought to reassure regional partners, especially GREEN. RED state-controlled media declared the recent arms sales from BLUE to GREEN and other nations in the region as an "unfair effort" to contain RED's ambitions to dominate. RED citizens are clamoring social media, calling for a "humbling" of BLUE forces in the region.

Since 2017, the BLUE Navy has adapted its fleet design to meet future challenges. Through a combination of doctrinal and materiel development, BLUE has transitioned from a "platform-centric" to a "fleet-centric" force. Therefore, air and surface kill chains are highly resilient, consisting of networked ships, aircraft, weapons, and unmanned systems.

2. Closing the Straits of Hormuz

In January 2020, GREEN's merchant marine was denied passage by RED IRC ships, as specified in the United Nations Convention on the Law of the Sea (UNCLOS). For two weeks, RED denied others from entering the gulf as well.

3. Political Situation

RED political will to start an <u>actual conflict imminently is considered high</u>. RED may perceive the upcoming BLUE basing and GREEN military acquisitions as a potential threat, giving them a real incentive to keep the strait closed to punish other regional BLUE allies from selling and shipping oil. BLUE is an ally of GREEN, and treaty-bound to defend it. Further, it is an election year for BLUE's president, and failure to show resolve during this quickly escalating foreign policy crisis may cost him reelection.

4. Military Activity

BLUE satellite imagery shows a RED combined IRC and conventional navy force used "previously planned" exercise to push forward with closing the strait. All RED Navy forces near Bandar Abbas are fully manned with its highest readiness and deployment levels not seen in years. RED's naval regions have started to sortie ships and submarines, and that they have increased maritime patrols on both sides of the strait. RED has also started mobilizing its SAM assets, transferring anti-ship cruise missiles (ASCM) from hardened storage facilities to working magazines. Air assets are conducting ISR, while tactical aircraft are flying sorties to increase proficiency.

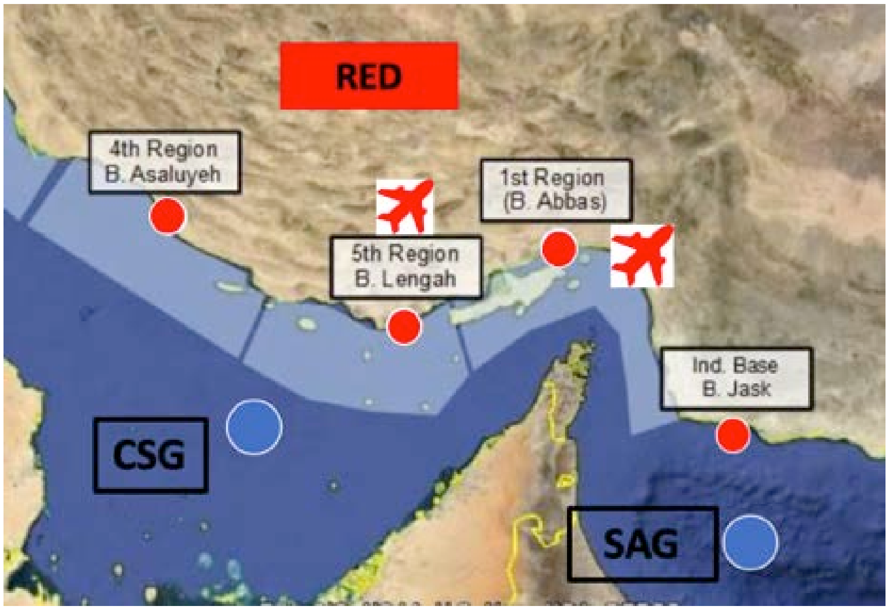


Figure 16. Intelligence summary

BLUE is determined to support GREEN, and keeping the straits open. They have forces already in the gulf, but also deployed forces to respond to RED threats and deter hostile action. Commander Carrier Strike Group (CSG) ONE is in the Arabian Sea conducting exercises and is scheduled to arrive at SOH within a day. A surge-ready CSG in the Baltic Sea is capable, but cannot arrive on station for another week. Therefore, Commander, US Fifth Fleet (C5F) must rely on forces already in the vicinity. With the strait already closed, BLUE must respond with forces on hand until

D. STRAIT OF HORMUZ BLUE SCENE SETTER

1. Purpose

LITMUS and jCORE – Strait of Hormuz explore Fleet Design in naval warfare at the operational level. Like any wargame, they are designed to capture the human elements of warfare. Therefore, players will be required to make decisions in the face of uncertainty.

2. Game Design

LITMUS and jCORE are closed wargames, meaning that opposing teams will play on opposite sides of room. Each team will have knowledge of:

1. Events that led to conflict (scenario)
2. Objectives (mission goals), provided by leadership and higher headquarters
3. Own force composition and capabilities
4. Capabilities of possible enemy platforms

Each team will have incomplete knowledge of:

1. The true enemy objective
2. The enemy force composition

Teams will have to make decisions based on:

1. The intelligence, surveillance, and reconnaissance (ISR) plans they develop
2. Technical and intelligence injects

Adjudication will rest with the student researcher, based on wargaming experience, analysis and knowledge of the combat modeling tools.

Blue Commander's intent

Maintain a presence in the gulf. Keep the SOH open to commercial traffic. Stay out of neutral country's TTW. Attempt to minimize any acts which could be construed as hostile intent by country RED.

- If fired upon, exercise self-defense with proportionality. Trip wires such as FAC (or larger vessel) fire-control radar may be considered hostile intent.

- Expect Red naval forces to probe defenses and TTP for FAC/FIAC. Do not pursue or react unless an individual vessel maneuvers within 500 feet, or a group within ¼ NM.

- Establish continuous search of vital area and of any Red naval forces underway. All assets are to bear the responsibility of, and share the duties of search for time in Arabian Gulf.

- Any interaction shall be professional, and IAW UNCLOS.

- Maintain open SLOC from country Green to Gulf of Oman (GOO) for white shipping.

- All asset Link-capable shall immediately share information up the appropriate CoC. Should there be a Link or GPS-denied environment, follow appropriate guidance.

Posture: Rd / Ti.

Mission Goals

Each force has is a set of desired outcomes from game play. Below are the details in order to achieve a "win" or "loss". Should one side not definitively achieve the requisite number of conditions, then it will each win condition is worth one point. The side with the most points wins, while a tie is a "détente".

BLUE Forces Goals

A. In order to win, BLUE must achieve three of the following while adhering to other specifics:

- Achieve a kill ratio equal of 5:1 for all assets

- Maintain a “presence” and ability to defend SOH for over 50% of game play
- Achieve air superiority (8:1 ratio)
- Neutralize military capability from Bandar Abbas
- Do not commence hostilities, but adhere to inherent right to self-defense and pre-planned responses

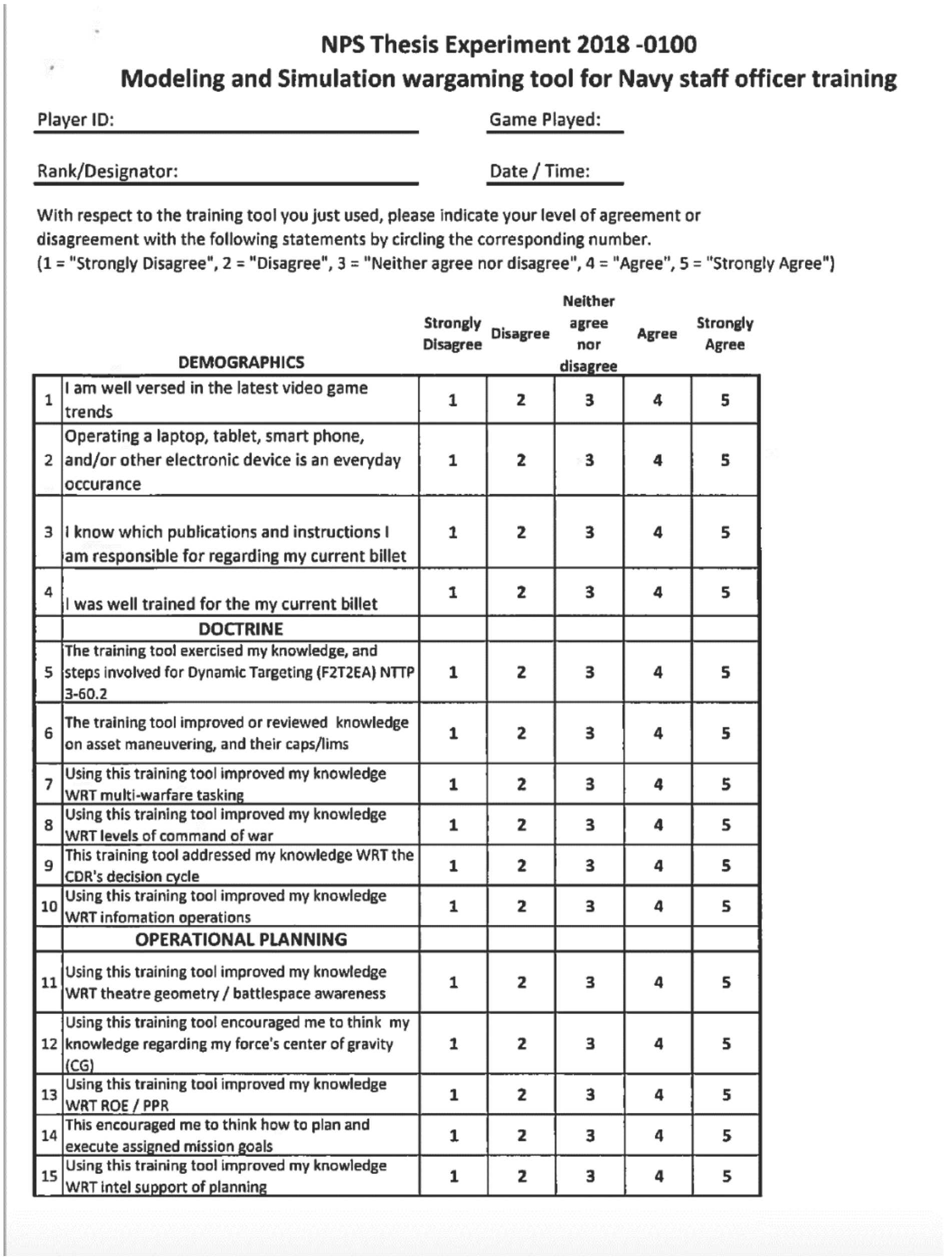
B. BLUE loses if any of the following conditions exist, ending game play:

- Blue CVN is sunk
- An entire SAG / CSG is not capable of conducting offensive operations at the end of game play
- BLUE units seek shelter from GREEN or neutral country in gulf (goes inside territorial waters to avoid conflict)
- BLUE player chooses wrong defense capability vs threat over 30% of time

THIS PAGE INTENTIONALLY LEFT BLANK

APPENDIX D. SCENARIOS

The IRB-approved survey was created in excel. The survey was briefed to the experiment audience so that questions were not erroneous filled in since we learned in Phase I that there could be multiple ways of interpreting the questions.



NPS Thesis Experiment 2018 -0100

Modeling and Simulation wargaming tool for Navy staff officer training

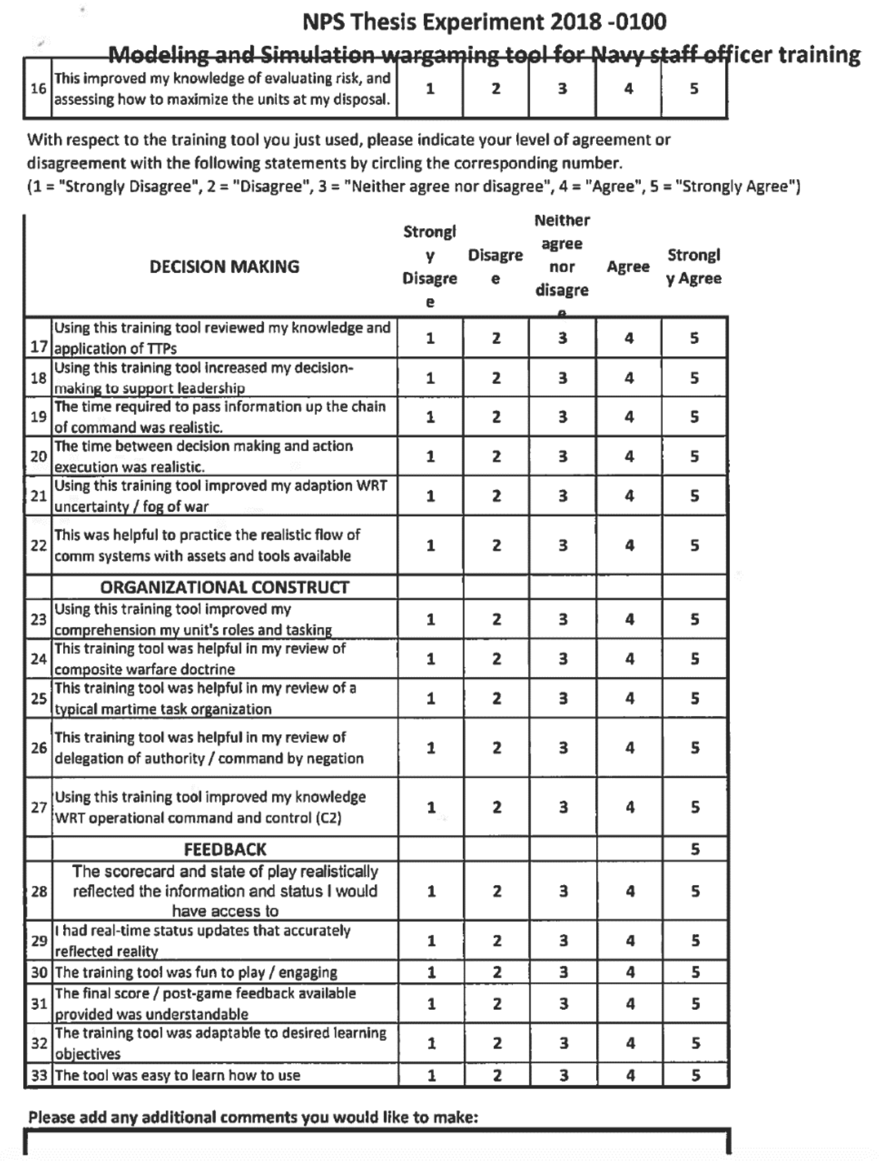
Player ID: ____________________ Game Played: ________

Rank/Designator: ____________________ Date / Time: ________

With respect to the training tool you just used, please indicate your level of agreement or disagreement with the following statements by circling the corresponding number.
(1 = "Strongly Disagree", 2 = "Disagree", 3 = "Neither agree nor disagree", 4 = "Agree", 5 = "Strongly Agree")

	DEMOGRAPHICS	Strongly Disagree	Disagree	Neither agree nor disagree	Agree	Strongly Agree
1	I am well versed in the latest video game trends	1	2	3	4	5
2	Operating a laptop, tablet, smart phone, and/or other electronic device is an everyday occurance	1	2	3	4	5
3	I know which publications and instructions I am responsible for regarding my current billet	1	2	3	4	5
4	I was well trained for the my current billet	1	2	3	4	5
	DOCTRINE					
5	The training tool exercised my knowledge, and steps involved for Dynamic Targeting (F2T2EA) NTTP 3-60.2	1	2	3	4	5
6	The training tool improved or reviewed knowledge on asset maneuvering, and their caps/lims	1	2	3	4	5
7	Using this training tool improved my knowledge WRT multi-warfare tasking	1	2	3	4	5
8	Using this training tool improved my knowledge WRT levels of command of war	1	2	3	4	5
9	This training tool addressed my knowledge WRT the CDR's decision cycle	1	2	3	4	5
10	Using this training tool improved my knowledge WRT infomation operations	1	2	3	4	5
	OPERATIONAL PLANNING					
11	Using this training tool improved my knowledge WRT theatre geometry / battlespace awareness	1	2	3	4	5
12	Using this training tool encouraged me to think my knowledge regarding my force's center of gravity (CG)	1	2	3	4	5
13	Using this training tool improved my knowledge WRT ROE / PPR	1	2	3	4	5
14	This encouraged me to think how to plan and execute assigned mission goals	1	2	3	4	5
15	Using this training tool improved my knowledge WRT intel support of planning	1	2	3	4	5

Figure 17.   Experiment survey page 1 of 2



NPS Thesis Experiment 2018 -0100

Modeling and Simulation wargaming tool for Navy staff officer training

16	This improved my knowledge of evaluating risk, and assessing how to maximize the units at my disposal.	1	2	3	4	5

With respect to the training tool you just used, please indicate your level of agreement or disagreement with the following statements by circling the corresponding number.
(1 = "Strongly Disagree", 2 = "Disagree", 3 = "Neither agree nor disagree", 4 = "Agree", 5 = "Strongly Agree")

	DECISION MAKING	Strongly Disagree	Disagree	Neither agree nor disagree	Agree	Strongly Agree
17	Using this training tool reviewed my knowledge and application of TTPs	1	2	3	4	5
18	Using this training tool increased my decision-making to support leadership	1	2	3	4	5
19	The time required to pass information up the chain of command was realistic.	1	2	3	4	5
20	The time between decision making and action execution was realistic.	1	2	3	4	5
21	Using this training tool improved my adaption WRT uncertainty / fog of war	1	2	3	4	5
22	This was helpful to practice the realistic flow of comm systems with assets and tools available	1	2	3	4	5
	ORGANIZATIONAL CONSTRUCT					
23	Using this training tool improved my comprehension my unit's roles and tasking	1	2	3	4	5
24	This training tool was helpful in my review of composite warfare doctrine	1	2	3	4	5
25	This training tool was helpful in my review of a typical martime task organization	1	2	3	4	5
26	This training tool was helpful in my review of delegation of authority / command by negation	1	2	3	4	5
27	Using this training tool improved my knowledge WRT operational command and control (C2)	1	2	3	4	5
	FEEDBACK					5
28	The scorecard and state of play realistically reflected the information and status I would have access to	1	2	3	4	5
29	I had real-time status updates that accurately reflected reality	1	2	3	4	5
30	The training tool was fun to play / engaging	1	2	3	4	5
31	The final score / post-game feedback available provided was understandable	1	2	3	4	5
32	The training tool was adaptable to desired learning objectives	1	2	3	4	5
33	The tool was easy to learn how to use	1	2	3	4	5

Please add any additional comments you would like to make:

Figure 18. Experiment survey page 2 of 2

LIST OF REFERENCES

Bartle, R. (1996, June). Hearts, clubs, diamonds, spades: Players who suit MUDs. Retrieved from https://www.researchgate.net/publication/247190693_Hearts_clubs_diamonds_spades_Players_who_suit_MUDs

Bloom, B. S. (1956). *Taxonomy of educational objectives, handbook I: The Cognitive domain.* New York, NY: David McKay Co. Inc.

Csikszentmihalyi, M. (1991). *Flow: The psychology of optimal experience: Steps toward enhancing the quality of life*. New York, NY: Harper Collins Publishers.

Dede, C. (2012). *Interweaving assessments into immersive authentic simulations: Design strategies for diagnostic and instructional insights* (commissioned white paper). Retrieved from http://www.k12center.org/rsc/pdf/session4-dede-paper-tea2012.pdf

Department of Defense. (2018). *The national defense strategy of the United States of America*. Washington, DC: Secretary of Defense.

Department of the Navy. (2010). *Naval warfare* (NDP-1). Retrieved from https://dnnlgwick.blob.core.windows.net/portals/14/Courses/Maritime%20Staff%20Operators%20Course/NDP-1-Naval-Warfare-(Mar-2010)_Chapters2-3.pdf?sr=b&si=DNNFileManagerPolicy&sig=2lMMssNQ%2FLyl1Fipw3oHsaF%2FKqAPTuJt6iVyiLbwKkA%3D

Department of the Navy. (2013). *Navy planning* (NWP 5–01). Retrieved from http://dnnlgwick.blob.core.windows.net/portals/10/MAWS/5-01_(Dec_2013)_(NWP)-(Promulgated).pdf?sr=b&si=DNNFileManagerPolicy&sig=un5q%2FWUW21Qzq52MmQ7KMfD%2FhHMdj%2Frp1xJSur5TF58%3D

Dörner, R., Göbel, S., Effelsberg, W., & Wiemeyer, J. (2016). *Serious games*. New York, NY: Springer International.

Hwang, G. & Chang, H. (2011). A formative assessment-based mobile learning approach to improving the learning attitudes and achievements of students. *Computers & Education*, *56*(4), 1023–1031. Retrieved from https://doi.org/10.1016/j.compedu.2010.12.002

Iten, N., & Petko, D (2014). *Learning with serious games: Is fun playing the game a predictor of learning success?* Retrieved from https://doi.org/10.1111/bjet.12226.

Joint Chiefs of Staff. (2018). *Joint maritime operations* (JP 3-32). Retrieved from https://www.jcs.mil/Portals/36/Documents/Doctrine/pubs/jp3_32.pdf?ver=2019-03-14-144800-240

Korteling, J., Helsdingen, A., Sluimer, R., van Emmerik, M., & Kappé, B. (2011). *Transfer of gaming: Transfer of training in serious gaming*. (Report No. TNO-DV 2011 B142). Retrieved from http://files.goc.nl/files/pdf/Gaming/2011%20Gaming%20transfer_gaming.pdf.

Lazzaro, N. (2004) Why we play games: Four keys to more emotion in player experiences. Paper presented at the of Game Development Conference, San Jose, CA. Retrieved from https://archive.org/stream/GDC2004Lazzaro/GDC2004-Lazzaro_djvu.txt

Lee, C. (2018, November 27). Navy pushes live, virtual, constructive training. *National Defense Magazine*. Retrieved from http://www.nationaldefensemagazine.org/articles/2018/11/27/navy-pushes-live-virtual-constructive-training

Loh, C. S., Sheng, Y., & Ifenthaler, D. (2015). *Serious games analytics: Theoretical framework*. New York, NY: Springer International.

Marczewski, A. (2012 November 30*)*. Flow, player journey and employee satisfaction [blog]. Retrieved from https://www.gamified.uk/2012/11/30/flow-and-satisfaction/.

Mayer, I., Bekebrede, G., Harteveld, C., Warmelink, H., Zhou, Q., van Ruijven, T., Lo, J., Kortmann, R., & Wenzler, I. (2014). The research and evaluation of serious games: Toward a comprehensive methodology. *British Journal of Educational Technology 45*, 502–507. Retrieved from https://doi.org/10.1111/bjet.12067.

McDowell, P. (2016). Evaluation of Strike Group Defender as a training platform. Retrieved from https://calhoun.nps.edu/bitstream/handle/10945/51928/NPS-MV-16-004 corrected.pdf?sequence=3&isAllowed=y

Merrill, M.D. (2013). *First principles of instruction: Identifying and designing effective, efficient and engaging instruction*. Hoboken, NJ: Pfeiffer.

Mislevy, R., Geneva, H., Riconscente, M., Rutstein, D, & Ziker, C. (2017). Evidence-centered assessment design (pp. 19-24). In *Assessing model-based reasoning using evidence-centered design*. New York, NY: Springer. Retrieved from https://doi.org/10.1007/978-3-319-52246-3_3

Novak, E., & Johnson, T. (2012). *Assessment of student's emotions in game-based learning*. New York, NY: Springer. Retrieved from https://doi.org/10.1007/978-1-4614-3546-4_19

President of the United States. (2017). *The national security strategy of the United States of America*. Washington, DC: Author.

RDocumentation. (n.d.-a). cox.stuart.test. Retrieved January 17, 2019, from https://www.rdocumentation.org/packages/webr/versions/0.1.0/topics/cox.stuart.test

RDocumentation. (n.d.-b). exANOVA. Retrieved April 5, 2019, from https://www.rdocumentation.org/packages/ez/versions/3.0-1/topics/ezANOVA

RDocumentation. (n.d.-c). SIGN.test. Retrieved January 10, 2019, from https://www.rdocumentation.org/packages/PASWR/versions/1.1/topics/SIGN.test

Scheldrup, M. (2018). *Effects of Level of Automation on Training and Mental Model Formation in a Real-world Command and Control Task* (Doctoral dissertation). Retrieved from https://psychology.gmu.edu/defenses/1087

Tekinbaş, K. S., & Zimmerman, E. (2003). *Rules of play: Game design fundamentals*. Cambridge, MA: MIT Press.

U.S. Army. (2002). America's Army [computer software]. Washington, DC: Author.

U.S. Fleet Forces Command, U.S. Pacific Fleet Command. (2012). F*leet training continuum instruction* (USFF-CPFINST 3501). Retrieved from http://elearning.sabrewebhosting.com/supportfiles/pdfs/USFF-CPFINST%203501%203D.pdf

U.S. Fleet Forces Command, II Marine Expeditionary Forces. (2016). *Amphibious Ready Group Fleet Response Training Plan and Marine Expeditionary Unit Predeployment Training Program* (COMUSFLTFORCOM/II MEF INST 3502.1).

Van Staalduinen, J., & De Freitas, Sara. (2011). *A Game-Based Learning Framework: Linking Game Design and Learning Outcomes. Learning to Play: Exploring the Future of Education with Video Games*. New York, NY: Springer International.

Wiemeyer, J & Hardy, S. (2013). Serious games and motor learning: Concepts, evidence, technology (pp. 197–208). In *Serious games and virtual worlds in education, professional development, and healthcare*, K. Bredl & W. Bosche (eds.). Hershey, PA: IGI Global. Retrieved from https://doi.org/10.4018/978-1-4666-3673-6.ch013

THIS PAGE INTENTIONALLY LEFT BLANK

INITIAL DISTRIBUTION LIST

1. Defense Technical Information Center
   Ft. Belvoir, Virginia

2. Dudley Knox Library
   Naval Postgraduate School
   Monterey, California