Calhoun: The NPS Institutional Archive

DSpace Repository

Acquisition Research Program | Acquisition Research Symposium

2017-10

An Empirical Study on Content Analysis Use in Test and Evaluation Deficiency Report Analysis

Holness, Karen S.; Khan, Rabia; Parker, Gary W.

Monterey, California. Naval Postgraduate School

http://hdl.handle.net/10945/58908

This publication is a work of the U.S. Government as defined in Title 17, United States Code, Section 101. Copyright protection is not available for this work in the United States.

Downloaded from NPS Archive: Calhoun

http://www.nps.edu/library

Calhoun is the Naval Postgraduate School's public access digital repository for research materials and institutional publications created by the NPS community. Calhoun is named for Professor of Mathematics Guy K. Calhoun, NPS's first appointed -- and published -- scholarly author.

Dudley Knox Library / Naval Postgraduate School
411 Dyer Road / 1 University Circle
Monterey, California USA 93943

NPS-TE-18-005

ACQUISITION RESEARCH PROGRAM SPONSORED REPORT SERIES

An Empirical Study on Content Analysis Use in Test and Evaluation Deficiency Report Analysis

11 October 2017

Dr. Karen Holness, Assistant Professor

Rabia Khan, Faculty Associate Research

Gary W. Parker, Faculty Associate Research

Department of Systems Engineering

Naval Postgraduate School

Approved for public release; distribution is unlimited.

Prepared for the Naval Postgraduate School, Monterey, CA 93943.

ACQUISITION RESEARCH PROGRAM
GRADUATE SCHOOL OF BUSINESS & PUBLIC POLICY
NAVAL POSTGRADUATE SCHOOL

The research presented in this report was supported by the Acquisition Research Program of the Graduate School of Business & Public Policy at the Naval Postgraduate School.

To request defense acquisition research, to become a research sponsor, or to print additional copies of reports, please contact any of the staff listed on the Acquisition Research Program website (www.acquisitionresearch.net).

ACQUISITION RESEARCH PROGRAM
GRADUATE SCHOOL OF BUSINESS & PUBLIC POLICY
NAVAL POSTGRADUATE SCHOOL

Abstract

This research investigated strategies and heuristics used to prioritize system deficiencies identified during test and evaluation. Five participants were recruited to participate in this laboratory study, and were assigned to an experiment condition either with or without content analysis training. Content analysis is a well-known methodology for identifying patterns and themes in qualitative datasets. In either experiment condition, subjects were asked to (1) classify a set of flight simulator deficiencies, (2) develop a deficiency resolution priority order using those classifications, and (3) complete a set of questionnaires regarding the completion of these tasks and demographic information. Across the five subjects, there was fairly high variability in the strategies and methods used. Therefore, the impact of the content analysis training was inconclusive. However, the variety of observed approaches warrants future research, specifically into the use of multiple categorization schemes when deciding upon a deficiency resolution priority order.

Keywords: content analysis, categorization, test and evaluation, deficiency reports, systems engineering

THIS PAGE INTENTIONALLY LEFT BLANK

Acknowledgments

The authors would like to acknowledge the support received from the Acquisition Research Program of the Graduate School of Business & Public Policy at the Naval Postgraduate School, without which this research study could not have been conducted. The authors also acknowledge the contributions of the anonymous study participants from the Naval Postgraduate School who, without compensation, volunteered their time as research subjects. Their participation will add to the body of knowledge and contribute towards improving the future education and training of systems engineering, acquisition, and test and evaluation personnel.

THIS PAGE INTENTIONALLY LEFT BLANK

About the Authors

Karen Holness, PhD

Karen Holness, PhD
Assistant Professor
Department of Systems Engineering
Naval Postgraduate School
Monterey, CA 93943-5194
Tel: (831) 656-2631
Fax: (831) 656-3840
E-mail: kholness@nps.edu

Rabia Khan

Rabia Khan
Faculty Associate for Research
Department of Systems Engineering
Naval Postgraduate School
Monterey, CA 93943-5194
Tel: (831) 656-2983
Fax: (831) 656-3840
E-mail: rhkhan@nps.edu

Gary W. Parker

Gary W. Parker
Faculty Associate for Research
Department of Systems Engineering
Naval Postgraduate School
Monterey, CA 93943-5194
Tel: (831) 656-7845
Fax: (831) 656-3840
E-mail: gwparker@nps.edu

THIS PAGE INTENTIONALLY LEFT BLANK

NPS-TE-18-005

ACQUISITION RESEARCH PROGRAM SPONSORED REPORT SERIES

An Empirical Study on Content Analysis Use in Test and Evaluation Deficiency Report Analysis

11 October 2017

Dr. Karen Holness, Assistant Professor

Rabia Khan, Faculty Associate Research

Gary W. Parker, Faculty Associate Research

Department of Systems Engineering

Naval Postgraduate School

Disclaimer: The views represented in this report are those of the authors and do not reflect the official policy position of the Navy, the Department of Defense, or the federal government.

THIS PAGE INTENTIONALLY LEFT BLANK

Table of Contents

THIS PAGE INTENTIONALLY LEFT BLANK

List of Figures

THIS PAGE INTENTIONALLY LEFT BLANK

List of Tables

THIS PAGE INTENTIONALLY LEFT BLANK

Executive Summary

The purpose of this study was to investigate the types of data evaluation, decision making, and planning strategies used when analyzing primarily qualitative test and evaluation (T&E) data. When documenting an observed deficiency, it is important to provide sufficient detail on what happened and provide an assessment of the deficiency's severity and implications. This assessment typically starts with a judgment of the system's ability to meet its operational and/or maintenance requirements in light of this failure. Following an investigation into the failure's root cause, there is a subsequent assessment of what it might take to fix it, what should be done to address it, and corresponding impacts to program cost and schedule. These assessments generate qualitative data that are leveraged to determine the order in which to work on, and resolve the deficiencies.

This research addressed one primary question: How can technical decision makers use patterns and themes in T&E data to prioritize the correction of system deficiencies discovered during test events? The content analysis methodology was used as the guiding framework for identifying patterns and themes.

This study was designed as a laboratory experiment. As human subjects research, the experimental protocols and materials were approved for use by the Naval Postgraduate School (NPS) Institutional Review Board (IRB) prior to the start of the experiment. The primary target population for this research was current NPS resident systems engineering (SE) students enrolled in either the 580, 308, or 581 (PhD program) curricula. Additional students were also recruited from the following curricula: 570—Naval/Mechanical Engineering (Total Ship Systems Engineering track); 816—Systems Acquisition Management; and 399—Modeling, Virtual Environments, & Simulation. A pilot study was conducted with one volunteer. Only four students volunteered to participate in the main experiment. For the final analysis, the results from the pilot study subject were included in the dataset, bringing the total number of participants to five.

Study participants were assigned to one of two experiment conditions where they either: (a) received a training session about content analysis and how to find patterns and themes using this method, or (b) received no training. In either condition, each participant was asked to categorize a list of deficiencies that were already assigned a technical priority by test personnel. Then, using the categories they created, subjects were asked to prioritize the deficiencies for resolution. Finally, using a series of questionnaires, subjects were asked to explain the thought processes they used to accomplish these tasks, provide ratings of task difficulty, ratings of impact and value of doing classifications, and demographic information.

Four out of five subjects (2, 3, 4, 5) created categories with an inherent or defined hierarchy within the categories themselves. The fifth subject's categories had no apparent hierarchy. Four out of five subjects (1, 2, 3, 5) created categories that described characteristics of the system's configuration (e.g., hardware, software, data, instructor, etc.) in order to group similar issues. Two out of five subjects (1, 4) used one category scheme to group similar issues, judge issue severity, and come up with a resolution priority order. These individuals reported the highest mental demand and temporal demand scores for this task. The remaining three subjects (2, 3, 5) used two category schemes. However, two of these subjects (2, 5) only used one of the schemes for the actual prioritization task. These subjects (2, 5) rated the impact of the categories the lowest in the questionnaire. Subject 3 was the only person to use the test personnel prioritization values as one of their schemes.

There were three key hypotheses for this study. The first hypothesis for this study was that subjects in the content analysis training condition would produce more well-defined categories than those in the non-training condition. No firm conclusion could be made regarding any training impact on the types of categories created or the number of category schemes used.

The second hypothesis for this study was that the perceived difficulty of the categorization and prioritization tasks (i.e., frustration level, mental and temporal demand, etc.) would be higher for those subjects in the non-training condition.

Based on the workload assessment results, no training impact was observed. The determining factors of perceived difficulty were the types of categories created and the number of category schemes used.

The third hypothesis for this study was that participants would leverage the issue prioritization assigned by the test personnel in order to come up with a resolution priority order. This strategy was expected by all participants, regardless of training condition. However, only one subject actually used the test personnel categorizations. The remaining four subjects created their own criteria to judge each issue's technical priority in order to sort them for resolution.

Based on these results, future research is needed to identify a categorization and prioritization scheme that produces consistent results across personnel from a variety of backgrounds. To start, it would be worth repeating this study, but provide incentives to increase volunteer enrollment. Once a consistent scheme is identified, further investigation to develop software tools and/or training for workforce development would be logical next steps.

THIS PAGE INTENTIONALLY LEFT BLANK

1.0 Introduction

Like other data analysis efforts within a typical Department of Defense (DoD) acquisition program, test and evaluation (T&E) data analysis efforts are generally constrained by program cost, schedule, and resource availability. However, the choice of analysis methodology also impacts the quality and reliability of the data analysis results. Government and contractor engineers who work with T&E data come from a variety of backgrounds, and have their own intuitive approaches to evaluating data. Therefore, the analysis of T&E data is further impacted by the mental models, heuristics, and biases inherent to each engineer working on the same dataset.

Holness (2016) described potential research in the use of content analysis in various systems engineering (SE) activities, including the Integration, Verification, and Validation processes of which T&E is a part. As an empirical study, this research investigated the types of data evaluation strategies, and corresponding decision making and planning strategies, used when analyzing primarily qualitative T&E data and leveraging a content analysis framework.

This research addressed one primary question: How can technical decision makers use patterns and themes in T&E data to prioritize the correction of system deficiencies discovered during test events? The following were the research objectives for this study:

a. Investigate the strategies and heuristics used by decision makers to
    i. identify patterns and themes in T&E datasets
    ii. use those patterns and themes to classify the deficiencies into categories
    iii. use those categories to prioritize deficiencies for resolution
b. Investigate the perceived level of effort and value of classifying data into categories

Throughout this report, variations on the terms *deficiency*, *discrepancy*, *anomaly*, *issue*, *problem*, *failure*, and *fault* are considered synonymous and are used interchangably.

THIS PAGE INTENTIONALLY LEFT BLANK

2.0 Literature Review

In support of either system verification or validation activities, the standard process for conducting a test and evaluation event involves adherence to a pre-established T&E plan with an approved set of test procedures. After executing the test procedures and recording the results, there is a need to analyze and resolve observed anomalies using some form of quality assurance process, to determine compliance with established requirements (International Council on Systems Engineering [INCOSE], 2015).

Kossiakoff, Sweet, Seymour, and Biemer (2011) state that the cause of discrepancies is not always obvious, since they can result from any number of factors, including issues with "(1) test equipment, (2) test procedures, (3) test execution, (4) test analysis, (5) the system under test, or (6) occasionally, to an excessively stringent performance requirement" (p.467). Wasson (2006) includes additional issues, like test environment and human error. He also states that when test failures occur, a discrepancy or deficient report (DR) is written, the significance of the problem on the system under test and the test plan needs to be determined, and the source of the failure must be isolated.

When documenting an observed deficiency, it is important to provide sufficient detail on what happened and provide an assessment of the deficiency's severity and implications. This assessment typically starts with a judgment of the system's ability to meet its operational and/or maintenance requirements in light of this failure. The most common way to do this uses a pre-determined classification scheme. For example, Kenett and Baker (2010) describe six generic severity classes for software, each with a corresponding generic definition: catastrophic, severe, moderate, minor, cosmetic, and comment. For example, *minor* is defined as when "things fail under very unusual circumstances, and recover pretty much by themselves. Users don't need to install any work-arounds, and performance impact is tolerable" (p. 196). Providing a descriptor for each severity class is important to support consistent use across developers and testers.

As shown in Figure 1, the sample Deficiency Report (DR) summary format, originally from the *Memorandum of Agreement (MOA) on Multi-Service Operational Test and Evaluation and Operational Suitability Terminology and Definitions* (2010) and shown in the DoD (2012) *Test and Evaluation Management Guide*, includes a column for deficiency description and an additional column for remarks. The deficiency, shown in Figure 1, was classified as minor. The written text and deficiency codes in discrepancy descriptions are qualitative and quantitative data that are also evaluated by the systems engineering team members to determine the best way to resolve the deficiencies.

Equip Nomen	Report I.D.	Report Date	Type of Deficiency	Deficiency Description	Cog. Agency	Closure Code	Action Ref	Remarks	Status	Date Information		
										Action AC CLO Date	Test for CLO Date	Last Update
	A		B		C	D						
AN/TCY-38 CNCE, ETC.	EPR 101-41.11-23001-YC-20-JFT, ETC.		INFO. MINOR. OPERATIONAL, ETC.	SHORT TITLE, PART NO, SUBASSEMBLY, ETC. PLUS PROGRAM EXAMPLES 1. OX-34 INVERTERS FAILED 2. SOFTWARE FLT-8 (E7R31) (DIAG) TRAINING PROBLEM WHEN TTY ON LINE. 3. YDU 8 CARD FAILURE	GTE, ESO, RCA, ETC.	NEEDHAM, FORT HUACHUCA, ETC.	FM-MS-404, ESD LTR 18 MAR 79	DEPOT REPAIR/REPLACE. TAPE PATCH DUE BY 24 AUG 79. SEE ECP AK-000, ETC.				

A. SERVICE UNIQUE REPORT NUMBER, i.e., EPR KH-41
B. TERMS LIKE "MAJOR," "MINOR," ETC.
C. WHERE THE CORRECTIVE ACTIONS WILL TAKE PLACE
D. PROBLEM REPORT #, DATE OF LETTER SENT TO AGENCY, ETC.

Figure 1. Sample Deficiency Report Summary. Source: MOA (2006).

In another DR summary example, the Naval Air Warfare Center Training Systems Division (NAWCTSD) uses a format that includes ample space for both a deficiency description and corrective action recommendation. It also includes a numerical deficiency category scale for any hardware, software, or process issue. As described on the NAWCTSD website (2017), "A Part I (critical), Part I* (safety/critical), Part II (major), or Part III (minor) DR classification shall be assigned to each deficiency."

Following an investigation into the failure's root cause, there is a subsequent assessment of what it might take to fix it, what should be done to address it, and

corresponding impacts to program cost and schedule. The order in which to work on the deficiencies is also determined. As stated in the DoD's (2012) T&E *Management Guide*, "A comprehensive and repeatable deficiency reporting process should be used throughout the acquisition process to report, evaluate, and track system deficiencies and to provide the impetus for corrective actions that improve performance to desired levels" (p.26).

Using the NAWCTSD categories as an example, it is clear that Part I and Part I* deficiencies must be addressed first, since they are critical and impact safety or mission execution. However, the Part II and Part III DRs must be reviewed for some order of precedence to be resolved and potentially retested by the test engineers. Depending on the size of the system, the number of DRs that need to be prioritized for resolution can vary from a few to many. The objective of this research study is to investigate different ways that system issues with assigned deficiency classifications are prioritized for resolution. Of particular interest is the creation and use of additional classification categories to complete this prioritization task.

2.1 Perspectives on Fault Classification and Prioritization

One of the most well-known classification schemes for software development and testing was introduced in 1992 by Chillarege, Bhandari, Chaar, Halliday, Moebus, Ray, and Wong. To give feedback to different developers, orthogonal defect classification (ODC) categorizes "defect types" discovered in the software development process and "defect triggers" from the software verification process. These categories are meant to be more semantic, as opposed to subjective and opinion based, in order to capture the variations that occur across organizations involved in the various design, code, and test activities. Leszak, Perry, and Stoll (2002) developed a software root cause analysis approach for network elements that is similar to ODC. However, their approach was used at the end of the development process, to produce retrospective feedback.

Chillarege et al. (1992) stated that "one of the pitfalls in classifying defects is that it is a human process, and is subject to the usual problems of human error,

confusion, and general distaste if the use of the data is not well understood" (p. 945). Leveraging ODC, Henningsson and Wohlin (2004) empirically investigated whether a group of independent software engineers could create the same classifications for a set of software faults as the actual developers, using only the fault descriptions. Eight subjects, with at least a master's degree in a software-related field, with or without industry experience, participated in this study on a voluntary basis. Given a list of fault descriptions, a classification scheme based on ODC, and forms to enter their responses, subjects completed the classification task. Subjects also completed a questionnaire about their confidence in their classification and completing classification tasks in general. The classification results were analyzed using a kappa statistic. Pairs of subject responses were compared using this method, with kappa values determined for all 28 pairs. The majority of the pairs, 27 out of 28, showed poor or fair agreement. However, the questionnaire results indicated a 3.51 out of 5 rating of confidence in the correctness of the fault classifications. The authors attributed these results to issues with the descriptions of the faults, the classification scheme, the background of the subjects, and their level of experience with the software used for this study.

Kamongi, Kotikela, Gomathisankaran, and Krishna (2013) added a weighted average mean to common vulnerability scoring system (CVSS) metrics used for assessing software system vulnerabilities. Coupled with the entered vulnerability scores for attributes such as "access complexity" or "confidentiality impact," the additional weighted average allows analysts to rank order identified attach paths to further prioritize security threats found in any software system or cloud application.

Specific to identifying, ranking, and prioritizing software code anomalies (aka code smells) that can lead to software architecture degradation, Arcoverde, Guimarães, Macía, Garcia, and Cai (2013) proposed four heuristics to aid developers: change-density, error-density, anomaly density, and architecture role. The authors identified anomalies within four software systems and scored them using the proposed heuristics by determining attribute counts within each heuristic. For example, for error-density, the number of errors resolved by changing certain code elements were counted; the higher the number of errors, the higher the

prioritization ranking. To determine the effectiveness of these heuristics, these scores were compared to a "ground-truth," rank-ordered, top-ten list of code element error sources independently produced by subject matter experts of these software systems. The similiarities in the scores varied significantly by software program, and the available data for the identified anomalies. The anomaly density heuristic was the only one applied to all four software data sets; all of the others were used for only three. However, the majority of similarity overlap scores exceeded their threshold value of 45%, indicating their effectiveness.

Outside of the software realm, the most well-known engineering-wide fault classification and prioritization method is failure mode and effects analysis (FMEA) with a calculated risk priority number, or failure modes, effects, and criticality analysis (FMECA), which has an additional calculated criticality value. When evaluating design alternatives, and conducting any verification and validation (V&V) activities during a typical SE conceptual or preliminary design phases, Jackson (2009) advocates for the use of this methodology to rank and prioritize faults, and implement any critical actions that are identified.

Other methods such as analytic hierarchy process (AHP) and multi-criteria decision analysis can also be used. Linkov, Satterstrom, and Fenton (2009) solicited judgments on prioritization criteria for capability gaps identified through the Joint Capabilities Integration & Development System (JCIDS) process using an online survey. Twenty-one participants from across the U.S. Army, Marine, Navy, Air Force, Coast Guard, and Special Operations Command responded, completed criteria pairwise comparisons, and assigned a numerical value to each comparison. These results were used for the AHP model to complete a multi-criteria decision analysis. At the end of the survey, participants were also asked to prioritize a set of capability gaps by ranking each gap with a subjective "gut feeling" importance value on a scale from one to nine. When these ad hoc results were compared to the decision analysis results, the authors realized that "although the top few gaps are ranked highly by both methods, other rankings differ significantly" (p. 184).

Many other types of ad hoc prioritizations are commonly used. In one example, Tao and Simons (2012) created a methodology in their workplace to identify and rank issues or "shortfalls" in terminal radar approach operations for air traffic control. After reviewing relevant data documents and identifying improvement areas, four general shortfall categories were identified: Airspace, Procedures, Decision Making, and Communications. Subject matter experts then identified specific shortfalls, categorized and ranked them by "how well and to what extent addressing each shortfall could resolve one or more of the improvement areas. … Shortfalls that addressed multiple improvement areas were ranked higher than other shortfalls" (p. I3-3). Then, with a list of capabilities mapped to the shortfalls, a capability value score formula was created that captured "the number of shortfalls a capability addressed, how well a capability addressed each of its assigned shortfalls, the importance of each shortfall being addressed. … Each score was normalized to a scale of 1 to 4. A capability's value, feasibility, and dependency scores were summed up to determine a 'Total Score'" (pp I3-5 – I3-6).

The DoD (2012) T&E *Management Guide* states that for software testing, a best practice is to "tailor the priority classification categories with respect to the actual system domain in coordination with end users" (p. 175). However, as evidenced by this brief literature review, tailoring applies to any system configuration, with hardware, software, people, and a governing process. When considering system issues that have already been classified into a deficiency category with an inherent order of predecence, additional evaluations must be made to order or rank the items within these categories to determine a resolution order. Essentially, categorizing and prioritizing means that a person reviews and interprets the written text descriptions of the issues, and considers the existing deficiency classification, to gain a holistic understanding of the issue's scope. Then, each issue's scope is compared with that of the other issues under consideration, to gauge which is more critical than another. The criteria used to gauge will vary because they can include (1) assessments of the part of the system impacted, (2) what it means for that part, (3) what it means for the parts it interfaces with, (4) overall system performance, and (5) overall program performance. For this reason,

the level of detail within the text descriptions and the accuracy of this data are key to a successful priority ranking.

Wickens and Carswell (2012) note that “the comprehensibility of text depends on many factors, from the reader’s experience, knowledge, and mental models that drive expectations to the structuring of text” (p. 136). Our mental models are grounded in our past experiences, and they influence how we interpret information, and subsequently our decision making and response selections. The basic human information processing model details how we perceive information we take in from our senses and interpret that information using what we have stored in our working and long term memory. Other things we simulataneously pay attention to also influence these tasks. When it comes to classifying and prioritizing, we are inherently problem solving and planning, which includes diagnosing, troubleshooting, and predicting the next steps to resolve the issue. All of these tasks are known to include some form of pattern matching technique, and are affected by heuristics and biases such as the availability heurisitc and confirmation bias (Wickens & Carswell, 2012). As described by Lehto, Nah, and Yi (2012), judgment takes place when a person “rates or assigns values to attributes of the alternatives considered” (p. 193). Judgments made during decision making tasks can include the use of heuristics such as representativeness, availability, and anchoring-and-adjustment.

Across a variety of deficiency classification and prioritization tasks, this literature review emphasizes that some combination of calculated numerical scores and human judgment is the common approach. More importantly, there is variability in how best to tailor an approach for a specific work domain. Of particular interest in this research is the creation and use of additional classification categories to complete a prioritization task. With this emphasis of embedding classification within prioritization, a discussion about the fundamentals of content analysis as a categorization process for qualitative data is warranted.

As defined by Patton (2015), content analysis refers to “any qualitative data reduction and sense-making efforts that takes a volume of qualitative material and attempts to identify core consistencies and meanings. … The core meanings found

through content analysis are patterns and themes" (p. 541). Under this general definition falls various methods for gathering relevant text segments, searching for occurrences of specific data points, iteratively coding the data, clustering data, then analyzing the results of the clusters and subsequent classifications for meaning and conclusions. This is the fundamental approach for grounded theory, defined by Birks and Mills (2012) as "an approach to research that aims to produce a theory, grounded in the data, through the application of essential methods" (p. 179). Further analyses using descriptive and inferential statistics such as frequency counts, chi-square, percent agreement, alpha and kappa statistics are used to evaluate classification schemes and gauge their validity when used by multiple coders (Krippendorff, 2013; Miles & Huberman, 1994; Patton, 2015). When determining inter-rater agreement and reliabilty, the best statistic to use in a specific content analysis study basically depends on the coding scheme, the number of raters, and the number of categories.

The goal of this research was to empirically investigate strategies individuals use to prioritize a list of deficiencies for resolution, with or without prior knowledge of the content analysis methodology. The design of this study is described in the next chapter.

3.0 Methodology

All research design and execution activities were completed at the Naval Postgraduate School (NPS) by the authors of this report. As human subjects research, the experimental protocols and materials were approved for use by the NPS Institutional Review Board (IRB) prior to the start of the experiment. The test materials used in the experiment were

- Unclassified and non-proprietary
- Understandable by a typical NPS Engineering and/or Graduate School of Business and Public Policy student
- Designed to target a specific deficiency prioritization solution

3.1 Experiment Design

The research study was designed as a laboratory experiment, where study participants sat in front of a computer and performed reading and assessment tasks using files created in standard office software such as Microsoft Word and Excel and Adobe Acrobat.

The primary target population for this research was current NPS resident systems engineering (SE) students enrolled in either the 580, 308, or 581 (PhD program) curricula. Additional students were recruited from the following curricula: 570—Naval/Mechanical Engineering (Total Ship Systems Engineering track), 816—Systems Acquisition Management, and 399—Modeling, Virtual Environments & Simulation. No previous experience with T&E was required to participate, no incentives were given to recruit subjects, and no compensation was provided to the volunteers at completion of the experiment. An informed consent form was used that explained participation was completely optional and that all data collected would be anonymonized.

Study participants were assigned to one of two experiment conditions where they either (a) received a training session about content analysis and how to find patterns and themes using this method or (b) received no training. In either

condition, each participant was asked to categorize a list of deficiencies that were already assigned a technical priority by test personnel using the previously described NAWCTSD deficiency codes. Then, using the categories they created, subjects were asked to prioritize the deficiencies for resolution and explain the thought processes they used to accomplish these tasks. The study was designed to be completed within two hours, regardless of experiment condition.

There were three key hypotheses guiding this study. First, the subjects in the content analysis training condition were expected to produce more well-defined categories than those in the non-training condition. Ideally, the training would assist with their category identification and classification strategy. Second, the perceived difficulty of the categorization and prioritization tasks (i.e., frustration level, mental and temporal demand, etc.) would be higher for those subjects in the non-training condition. Third, participants were expected to leverage the issue prioritization assigned by the test personnel in order to come up with a resolution priority order. In other words, all of the Part II issues labeled by the test personnel would have higher resolution priority numbers than the Part III issues, regardless of the issue categories the subjects created on their own. This was the expected deficiency prioritization solution. This strategy was also expected by all participants, regardless of training condition.

No power analysis was performed to determine the sample size for this study. The expected number of participants was 10–20 SE department students, based on the approximately 45–50 eligible students in the 580, 308, and 581 curricula during the 2017 summer quarter. This number seemed reasonable, based on sample sizes reported in similar studies from the research literature. As described in the previous chapter of this report, Henningsson and Wohlin (2004) had eight participants, while Linkov et al. (2009) had 21 participants. In a policy capturing study reported by Lafond, Vallieres, Vachon, St Louis, and Tremblay (2015), 60 university students performed a radar contact classification task in naval air-defense scenario using a simulated combat control system (S-CCS) microworld. Finally, in the Cropp, Banks, and Elghali (2011) study, 30 industry professional reviewed hypothetical case studies and rated potential risks associated with each one.

3.2 Data Collection Method

All data collection took place in a dedicated office space within the systems engineering department, preconfigured to provide a quiet, secure environment with all the necessary equipment, reading material, and supplies to complete the experiment within the given timeframe. The research associates proctored the experiment and answered any questions the subjects asked.

A pilot study was conducted prior to the main experiment, using an appropriately modified version of the recruitment script to indicate the pilot test run. One to three SE department faculty, research staff, and summer interns were expected to be recruited via email to participate in the pilot study. However, only one person volunteered to participate in the timeframe allotted. After evaluating this person's data, no changes were made to the methodology or data collection process.

For the main experiment, student participants were recruited via email. A copy of the informed consent form was attached to the email, so potential participants could read it ahead of time and decide if they wished to participate in this study. In addition to email, visits to some of the 580 and 308 classrooms were made to advertise the availability of the study and promote responses to the email. Students were asked to contact the research associates listed in the email if interested in participating, and indicate a day and time that worked best with their schedule. Recruitment took place in July and August 2017, and data collection took place in the month of August. Only four students volunteered to participate.

At the beginning of each experiment session, subjects were first asked to sign the informed consent form. Then, they were given an overview of what they were expected to do. Those in the training condition were asked to review a PowerPoint file with an 18 minute narrated instructional brief on content analysis methodology before starting the main experiment task. All subjects were asked to complete the following tasks:

- Read the provided T&E deficiency report that described testing performed on a generic aircraft flight simulator system.
- Using an Excel spreadsheet, look for patterns and themes in the provided deficiencies and create categories to help them prioritize the issues for resolution.
- Create a prioritized deficiency list indicating the order they think the deficiencies should be resolved.
- Complete a demographics questionnaire about their backgrounds and T&E experience.
- Complete questionnaires that assessed
    a. the classification strategies they used,
    b. perceived classification task difficulty,
    c. the value they assigned to doing the classification task as part of deficiency prioritization, and
    d. the impact the categories had on the priority order.

The provided T&E deficiency report was both generic and realistic, describing tests conducted on the flight simulator and deficiencies discovered during testing. The deficiencies were defined as issues found in the simulator's hardware and software by test personnel who executed a set of approved simulator test procedures. The deficiency list provided in the T&E report contained 25 issues. For each issue, a brief description was provided, along with the deficiency priority assigned by the test personnel and the name of the organization primarily responsible for resolving the issue. All of the deficiencies were either a Part II or Part III deficiency, as defined by the NAWCTSD guidance described previously. Definitions of all of the NAWCTSD classifications were provided in the T&E report for each subject's reference. The complete deficiency list used in this study is shown in Appendix A of this report.

Subjects were asked to view themselves as a government systems engineer, read through the list of identified deficiencies, group them into relevant categories, and use those categories to prioritize the deficiencies for resolution. The subjects were specifically instructed via a hardcopy instruction sheet to assign each deficiency a unique priority number (i.e., two or more deficiencies could not be

assigned the same priority number). For the purposes of the study, subjects were instructed to assume the following:

- Both funding and personnel are available to work on all identified issues.
- All issues must be resolved within the next 1–2 months.
- A resolution for each issue can be either a fix, a workaround solution, or planned deferral of resolution until something else is obtained.

Subjects did not have to identify a course of action to resolve each issue. Rather, they were asked to assume that one would be created for each deficiency after the priority order for resolution is complete. Using the provided T&E report as a reference and working with the list of deficiencies in a Microsoft Excel spreadsheet, the subjects were asked to complete the categorization and prioritization task within one hour. A pen and paper was provided to each subject during the course of the study, should they have wanted to write notes to assist in completing the tasks.

At the end of the prioritization task, the research associate noted the subject's completion time, then gave each subject an additional 15 minutes to complete a series of questionnaires in a separate Excel spreadsheet. These questionnaires were designed to capture the subject's demographic information, classification strategies, perceptions of task difficulty, and perceptions of the value of doing classifications as part of deficiency prioritization. These questionnaires can be found in Appendices B, C, D, and E of this report.

On completion of the questionnaires, the research associate provided a short debriefing, then collected any notes the subjects may have taken. Subjects were allowed to read and leave with a copy of the debrief form at the conclusion of the two-hour experiment block.

3.3 Data Analysis Method

The research associates uploaded all individual subject data files to a secure NPS file server. All of the subject responses to both the categoriziation/prioritization exercise and the questionnaire were anonymized and aggregated into a master

Excel spreadsheet. For analysis purposes, the pilot study results were included in the final dataset, bringing the total number of participants to five.

The initial data analysis approach was to apply a content analysis approach to the qualitative data collected from the subjects and apply descriptive and inferential statistics to the quantitative data. The low number of subjects that responded to the recruitment campaign limited the usefulness of inferential statistics. Instead, only frequency counts, averages, standard deviations, and pairwise comparisons of the numerical data were performed.

4.0 Results Summary

The participants included one NPS employee and four NPS students. Two students were from the SE curriculum, and two were from the 816 System Acquisition Management curriculum. Two of the students were current active duty, and two were civilians.

Across all five subjects, the reported bachelor's degrees included communication studies, mechanical engineering, business management, and oceanography. The reported master's degrees included management, aerospace engineering, and national security and strategic studies. No subjects held a PhD in any field.

Only two subjects had prior experience evaluating T&E data, each reporting five and seven years of experience. Three of the five subjects were assigned to the content analysis training condition; two did not receive the training. Both subjects in the non-training condition took slightly more than an hour to complete the classification and prioritization task, as did one of the subjects who received the training. The other two subjects in the training condition took less than one hour to complete the task. Across the five subjects, the average time to complete these tasks was 58 minutes.

4.1 Categorization Results

Table 1 shows a sample of the results of the categorization exercise for the flight simulator Part II issues. The complete table of these results can be found in Appendix F. The results were grouped by training condition so as to highlight any substantial similarities and/or differences between the two subject groups.

Subjects 1 and 4 created one category scheme, while the remaining subjects created two category schemes. Subject 3 was the only person to incorporate the Test Personnel prioritizations into their categorization and prioritization scheme. Subjects 1 and 3, who were both in the training condition, had the most similar

hardware and software categorizations. Subjects 2 and 5 created categories related to specific types of hardware, software, and other system elements (e.g., instructor, procedure). Of particular interest is the fact that four out of five subjects created a scheme with an inherent or defined hierarchy. Even Subject 3, who used the Test Personnel issue priority values, assigned an order of precedence to the second category set: (1) Additional information required/Possible Part I, (2) Hardware functionality missing/Testing not completed, (3) Software bug functionality missing/Testing not completed, (4) Software bug, (5) Non-functional hardware deficiency.

Table 1. Sample Part II Deficiency Categorization Results

Issue #	Issue Title	Category Subject 1 (T)	Category Subject 3 (T)	Category Subject 5 (T)	Category Subject 2 (NT)	Category Subject 4 (NT)
6	Missing Battery Indicator	Hardware	Part II. Hardware functionality missing. Testing not completed.	Ancillary, Priority D	Physical component, Part III	Minor
7	Headset Mic Problem	Hardware	Part II. Additional information required on availability of workaround and what the contract specified. Potential to be a Part I.	Ancillary, Priority D	Interface, Part II	Major
8	Instructor Station– Screen capture software test incomplete	Hardware	Part II. Hardware functionality missing. Testing not completed.	Instructor, Priority B	Data capture, Part III	Minor
9	Digital Map malfunction	Simulation Software	Part II. Software bug.	Cockpit, Priority B	Procedure mismatch, Part II	Critical
13	Flap display not working	Hardware	Part II. Software bug.	Cockpit, Priority B	Procedure mismatch, Part I	Minor
15	Visual Scene– Time of Day mismatch	Simulation Software	Part II. Software bug.	Visual, Priority C	Visual system delta, Part III	Critical
22	Trainer automatic power shutdown did not work	Hardware	Part II. Software bug? Functionality missing. Testing not completed.	Ancillary, (safety), Priority A	Physical component, Part I*	Major

Key: (T) – Training condition; (NT) – Non-Training Condition

Also noteworthy is the fact that the two subjects assigned to the non-training condition seemed to leverage the NAWCTSD deficiency code definitions provided in the T&E report to create their categories. Table 2 shows a sample of the results of the categorization exercise for the flight simulator Part III issues. The complete table of results can also be found in Appendix F.

Table 2. Sample Part III Deficiency Categorization Results

Issue #	Issue Title	Category Subject 1 (T)	Category Subject 3 (T)	Category Subject 5 (T)	Category Subject 2 (NT)	Category Subject 4 (NT)
1	Coldstart media missing	Technical Software	Part III. Hardware functionality missing. Testing not completed.	Data, Priority A	Physical component, Part I	Critical
2	Can't play back recorded mission	Technical Software	Part III. Software bug.	Instructor, Priority A	Data capture, Part I	Major
4	Lighting system mismatch	Hardware	Part III. Hardware functionality missing. Testing not completed.	Cockpit, Priority D	Physical component, Part II	Minor
10	Ice Shedding/ Removal	Simulation Software	Part III. Software bug.	Visual, Priority C	Procedure mismatch, Part III	Major
11	Gross Weight	Simulation Software	Part III. Software bug.	Instructor, Priority B	Procedure mismatch, Part III	Critical
12	Engine Fire Extinguisher malfunction buttons	Hardware	Part III. Software bug.	Cockpit, Priority A	Procedure mismatch, Part I	Safety/critical
23	No audio captured in mission recording	Technical Software	Part III. Additional information required on availability of workaround and what the contract specified. Potential to be a Part I.	Instructor, Priority A	Data capture, Part I	Critical

Key: (T) – Training condition; (NT) – Non-Training Condition

The results from Tables 1 and 2 highlight the differences in approach to assign issues to the created categories. Given the aforementioned observations on the categorization strategies used by the test subjects, it can be seen that a heuristic used by the subjects to focus on high-level attributes of the system, perhaps as a way to manage and consolidate the data in a meaningful way. Utilizing the provided descriptions of each issue, and their own interpretations and mental model of each issue, each subject made a judgment of circumstance, scope and criticality. Despite the similarities in some of the category names, each person's working definition of these categories was different enough, such that the same issues were not all assigned to the same categories. It is difficult to tell what their categories would have looked like if they were specifically instructed to use the Test Personnel prioritizations. Based on these results, the impact of the content analysis training was inconclusive.

4.2 Prioritization Results

Each subject was asked to first categorize the issues, then prioritize the issues for remediation. Table 3 lists the assigned priority numbers for the Part II issues. The results were again grouped by subject condition so as to highlight any substantial similarities and/or differences between the two subject groups.

As directed by the experiment instructions, subjects were specifically asked to assign a unique priority number to each issue, without duplication of ranking (i.e., two or more deficiencies cannot be assigned the same priority number). Subjects 3, 4, and 5 used a 1–25 scale and assigned a unique resolution priority number to each issue. For the remaining two subjects,

- Subject 2 assigned all issues either a 1, 2 or 3. Even though this person created two category schemes, only the scheme with the inherent hierarchy (Part I, Part I*, Part II, Part III) was used for resolution prioritization. This resulted in multiple #1, #2 and #3 issues that require further prioritization within each of these subsets.
- Subject 1 used a scale dependent upon the number of issues in each category. In other words, the ten issues assigned to the "hardware" category were assigned resolution priority numbers 1–10. The twelve issues assigned

to the “simulation software” category were assigned resolution priority numbers 1–12. The three issues assigned to the “technical software” category were assigned resolution priority numbers 1–3. This strategy also resulted in multiple issues with the same resolution priority ranking that require further prioritization within each of these subsets.

There were 25 issues total: 11 Part II and 14 Part III. Had the priority from the test personnel been leveraged, it was expected that all of the Part II issues would appear within the top 11 rankings of the prioritization list. As shown in Table 3, this was the case for Subject 3. For Subjects 4 and 5, who also used a 1–25 scale, this was not the case at all, because of their interpretation of the issues and the categories they used. It is noteworthy that Subjects 3, 4, and 5 rated only one issue the same resolution priority number (Part III issue #20). The complete table of prioritization results for the Part III issues can be found in Appendix G.

Table 3. Part II Deficiency Prioritization Results

Issue #	Issue Title	Priority Assigned by Test Personnel	Priority for Subject 1 (T)	Priority for Subject 3 (T)	Priority for Subject 5 (T)	Priority for Subject 2 (NT)	Priority for Subject 4 (NT)
6	Missing Battery Indicator	II	6	2	24	2	24
7	Headset Mic Problem	II	2	1	22	2	9
8	Instructor Station–Screen capture software test incomplete	II	5	3	8	3	22
9	Digital Map malfunction	II	5	8	10	2	5
13	Flap display not working	II	4	5	9	1	21
15	Visual Scene–Time of Day mismatch	II	11	9	11	3	6
17	Incorrect weather depiction	II	4	10	18	3	8
18	Cross winds setup	II	2	6	4	1	17
19	Night FLIR not working	II	3	7	12	1	19
22	Trainer automatic power shutdown did not work	II	1	4	1	1	15
25	Weather visual scene and cockpit display mismatch	II	1	11	19	3	16

Priority Ranking Statistics

Table 4 summarizes the priority rankings assigned by the five subjects to each of the 25 deficiencies. Subjects 1 and 2 did not follow the instructions given to them to assign a unique priority ranking to each deficiency. Their responses are presented for completeness, but grayed out to indicate their incompatibility for use in any statistics. The average and standard deviation of the rankings by Subjects 3, 4, and 5 are shown at the right of the figure. A low standard deviation (like issues 20, 5, and 9) indicates closer agreement among the subjects than those issues with large standard deviations like Issues 6, 14, 7, and 3.

Table 4. Average and Standard Deviation of Issue Priority Ranking by Subjects 3, 4, and 5

	Issue Prioritization by Subject #					Subjects 3-5	
Issue #	Subj 1	Subj 2	Subj 3	Subj 4	Subj 5	Average	Std dev
1	3	1	15	3	2	6.7	7.234
2	1	1	16	10	6	10.7	5.033
3	9	3	23	18	3	14.7	10.408
4	7	2	13	23	25	20.3	6.429
5	8	3	14	14	15	14.3	0.577
6	6	2	2	24	24	16.7	12.702
7	2	2	1	9	22	10.7	10.599
8	5	3	3	22	8	11.0	9.849
9	5	2	8	5	10	7.7	2.517
10	7	3	22	11	16	16.3	5.508
11	6	3	18	4	13	11.7	7.095
12	3	1	17	2	7	8.7	7.638
13	4	1	5	21	9	11.7	8.327
14	12	3	21	1	21	14.3	11.547
15	11	3	9	6	11	8.7	2.517
16	10	3	25	25	17	22.3	4.619
17	4	3	10	8	18	12.0	5.292
18	2	1	6	17	4	9.0	7.000
19	3	1	7	19	12	12.7	6.028
20	8	3	20	20	20	20.0	0.000
21	10	3	24	12	23	19.7	6.658
22	1	1	4	15	1	6.7	7.371
23	2	1	12	7	5	8.0	3.606
24	9	3	19	13	14	15.3	3.215
25	1	3	11	16	19	15.3	4.041

Figure 2 shows a bar graph of the average ranking for each issue.

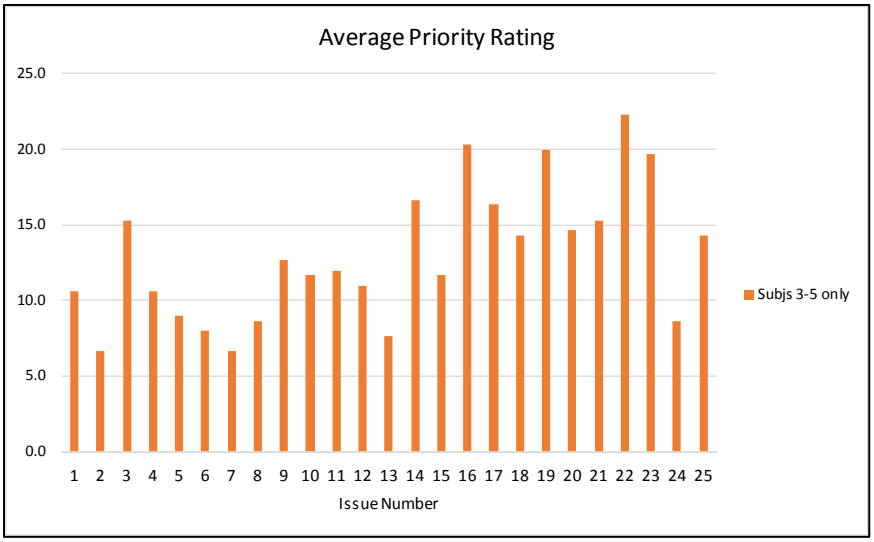

Figure 2. Average of Priority Ratings by Issue Number

Since the same average ranking could be obtained from different sets of widely differing data, it is instructive (given the small number of subjects) to do a pairwise comparison of rankings between subjects. Table 5 shows the ranking data again, with the right three columns showing the absolute value of the difference in ranking among the three possible pairs of subjects (i.e., Subjects 3 and 4, 3 and 5, and 4 and 5). Comparisons where the subjects agree have a distance of zero and are highlighted in green.

Lastly, Figure 3 shows graphically the spread of priority rankings for the 25 deficiencies between pairs of subjects. Such a graph highlights issues where there was close agreement (e.g., Issue 20) and wide disagreement such as Subjects 3 and 4 on Issues 10, 11, 12, 13, and 14.

Given these findings, no statistically significant differences/statistically significant differences were observed in the perceived value, between those in the training condition and those in the control condition. Once again, based on these results, the impact of the content analysis training was inconclusive.

Table 5. Pairwise Comparison of Ranking Spread by Issue

Issue Prioritization by Subject #								
Issue #	Subj 1	Subj 2	Subj 3	Subj 4	Subj 5	3-4 Spread	3-5 Spread	4-5 Spread
1	3	1	15	3	2	12	13	1
2	1	1	16	10	6	6	10	4
3	9	3	23	18	3	5	20	15
4	7	2	13	23	25	10	12	2
5	8	3	14	14	15	0	1	1
6	6	2	2	24	24	22	22	0
7	2	2	1	9	22	8	21	13
8	5	3	3	22	8	19	5	14
9	5	2	8	5	10	3	2	5
10	7	3	22	11	16	11	6	5
11	6	3	18	4	13	14	5	9
12	3	1	17	2	7	15	10	5
13	4	1	5	21	9	16	4	12
14	12	3	21	1	21	20	0	20
15	11	3	9	6	11	3	2	5
16	10	3	25	25	17	0	8	8
17	4	3	10	8	18	2	8	10
18	2	1	6	17	4	11	2	13
19	3	1	7	19	12	12	5	7
20	8	3	20	20	20	0	0	0
21	10	3	24	12	23	12	1	11
22	1	1	4	15	1	11	3	14
23	2	1	12	7	5	5	7	2
24	9	3	19	13	14	6	5	1
25	1	3	11	16	19	5	8	3

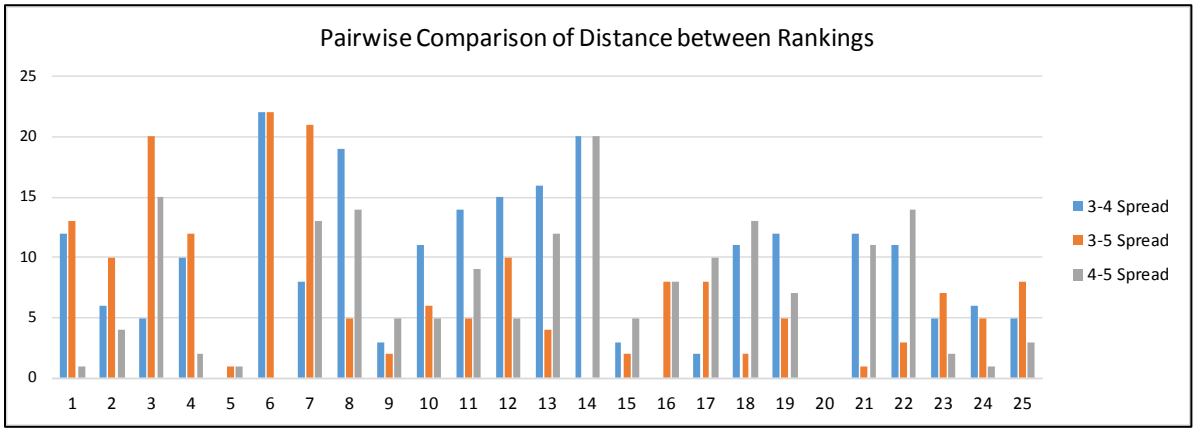

Figure 3. Pairwise Comparison of Distance Between Rankings by Issue

4.3 Classification Strategies Questionnaire Results

A content analysis on the classification questionnaire data was performed to study whether or not there were observable differences in strategies between the subjects in the training condition and the control condition. The content analysis was based on keywords used, with the objective to evaluate themes based on the number of common phrases or keywords found in the results.

The subjects described the rationale they used to create categories, assign a resolution priority number to each issue, and reconcile which category or priority number to assign an issue if more than one seemed to apply. As shown in Figure 4, the reported answers were summarized into seven types:

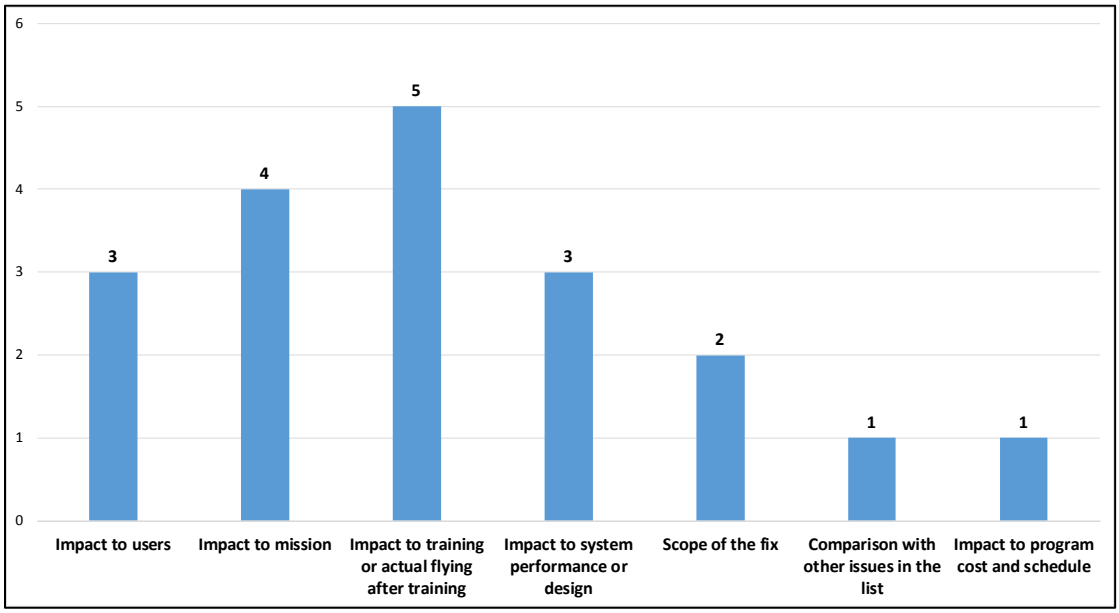

Figure 4. Counts of Reported Rationale

Impact to users, mission, training, or actual flying after training seem similar and could be consolidated. However, more detailed rationale is required to group them together. No noticeable differences between subjects in the training versus non-training condition were found. It is interesting to note that one subject specifically noted looking for “patterns of deficiencies” as a classification strategy. This person was not in the training condition, but did have a background in the T&E domain.

When asked if they leveraged anything from their previous training or experience, prior problem solving was a commonly cited theme across all subjects.

4.4 Workload Assessment Questionnaire Results

The results outlined in Table 6 are the subject responses to the workload assessment questionnaires. Subjects were asked to rate their perceived level of workload, on a number of factors, from a scale from 1 to 10, with "1" being the lowest score, or reflecting a "poor" level and "10" being the highest score, or a "good" level.

Table 6. Workload Assessment Questionnaire Results

	Mental	Temporal Demand	Performance	Effort	Frustration Level
Subject 2 (NT)	4	5	7	6	3
Subject 4 (NT)	9	10	8	10	9
Subject 1 (T)	9	9	4	6	4
Subject 3 (T)	8	2	6	8	3
Subject 5 (T)	6	6	4	5	3
Average Rating:	**7**	**6**	**6**	**7**	**4**

In general, subjects in the training condition rated the mental demand to be high, but the frustration level low. The high scores for Subject 4 in the non-training condition were attributed to the fact that this person was an international, non-native English speaking student who had no prior T&E experience. Subjects 2 and 3, who rated the lowest temporal demand, were the ones that took the longest to complete the task. Even though subjects were told they had up to one hour to complete the categorization and prioritization tasks, these subjects exceeded the allotted hour by 8 minutes and 14 minutes, respectively. The individuals who reported the highest mental demand and temporal demand scores for this task, Subjects 1 and 4, both used one category scheme to group similar issues, judge issue severity, and come up with a resolution priority order.

An interesting observation on the performance attribute is that those in the training condition rated their overall level of satisfaction with completing the tasks

lower than those in the non-training condition. Additional data is needed to determine an explanation.

4.5 Perceived Value Questionnaire Results

The subjects were asked to rate two factors: (1) the value of categorizing deficiencies before prioritizing them, and (2) the impact of categorizing on prioritization order. The subjects were asked to use a scale from 1 to 10, with "1" being the lowest score, or reflecting a "low" perceived value and "10" being the highest score, or a "high" perceived value. Table 7 summarizes these responses. No significant differences were observed between those in the training condition and the control condition on the value scores.

Table 7. Value and Impact of Categories Questionnaire Results

Subject	Value of Categories	Impact of Categories
Subject 2 (NT)	No score provided	No score provided
Subject 4 (NT)	10	10
Subject 1 (T)	10	10
Subject 3 (T)	6	8
Subject 5 (T)	8	1
Average Rating:	8	6

Subject 2 provided comments for both of these questions instead of a numerical score. In essence, this person described categorizing the deficiencies after prioritizing them, and stated the belief that mission impact is the most important consideration in prioritization. An interesting observation is that Subjects 2 and 5 used two category schemes: one with an inherent hierarchy and one specific to system characteristics. However, neither of them used the latter during the prioritization task. This also explains the low impact score provided by Subject 5. The remaining Subjects 1, 3, and 4 all used their categories to help them assign resolution priority numbers to the issues.

THIS PAGE INTENTIONALLY LEFT BLANK

5.0 Discussion

The goal of this study was to evaluate ways that deficiencies are classified into categories using available information and how the correction of the deficiencies is prioritized using those categories.

The first hypothesis for this study was that subjects in the content analysis training condition would produce more well-defined categories than those in the non-training condition. No firm conclusion could be made regarding any training impact on the types of categories created or the number of category schemes used.

The second hypothesis for this study was that the perceived difficulty of the categorization and prioritization tasks (i.e., frustration level, mental and temporal demand, etc.) would be higher for those subjects in the non-training condition. Based on the workload assessment results, no training impact was observed. The determining factors of perceived difficulty were the types of categories created and the number of category schemes used.

The third hypothesis for this study was that participants would leverage the issue prioritization assigned by the test personnel in order to come up with a resolution priority order. This strategy was expected by all participants, regardless of training condition. Only one subject actually used the test personnel categorizations. The remaining four subjects created their own criteria to judge each issue's technical priority in order to sort them for resolution. Only one subject explained why the issue priority assigned by the test personnel was not used. In this person's opinion, test personnel often do not have adequate training or operational experience as a system user to judge the criticality of issues identified during test. It should be noted that this bias was stated by a subject that self-reported no prior T&E experience.

All subjects realized a need to judge the severity of each issue using the information provided, and their own past experience with classification and prioritization, to come up with a resolution priority order. However, the strategies they used were very different, with a high degree of subjectivity in methodology

used. It was not possible to determine which interpretations and approaches were the most efficient in terms of time to complete and level of effort. There were no apparent correlations between educational background, prior T&E experience, and strategy used. With a greater number of study participants, more repetition in similar strategies might have been observed.

Future Research

Because of the small number of participants recruited in this study, it would be worth repeating, but with incentives provided to increase volunteer enrollment. The results of this study indicate that using both a technology-based and priority-based categorization scheme might produce more consistent results across subjects. It would be interesting to revise this research study to only investigate how subjects assign issues to pre-defined technology-based and priority-based categories provided to them. Another variation is to pre-assign issues to such categories, then ask subjects to create a resolution priority order. Finally, it seems worth investigating the preferences people have for resolution prioritization criteria. The results of this study indicate a preference for ordinal versus interval criteria and measurement scales.

The ultimate objective of further research in this topic is to generate a categorization and prioritization scheme that produces consistent results across personnel from a variety of backgrounds. Ideally, with a valid scheme, the only key differentiating factor between personnel would be their level of domain knowledge and T&E experience with a specific type of system. With such a scheme identified, further research to develop software tools and/or training for workforce development would be logical next steps.

6.0 References

Arcoverde, R., Guimarães, E., Macía, I., Garcia, A., & Cai, Y. (2013). Prioritization of code anomalies based on architecture sensitiveness. In *Proceedings of the 27th Brazilian Symposium on Software Engineering* (pp. 69–78). doi: 10.1109/SBES.2013.14

Birks, M., & Mills, J. (2012). *Grounded theory: A practical guide*. Thousand Oaks, CA: Sage.

Chillarege, R., Bhandari, I. S., Chaar, J. K., Halliday, M. J., Moebus, D. S., Ray, B. K., & Wong, M. (1992). Orthogonal defect classification: A concept for in-process measurements. *IEEE Transactions on Software Engineering, 18*(11), 943–956., doi: 10.1109/32.177364

Cropp, N., Banks, A., & Elghali, L. (2011). Expert decision making in a complex engineering environment: A comparison of the lens model, explanatory coherence, and matching heuristics. *Journal of Cognitive Engineering and Decision Making*, *5*(3), 255–276. doi: 10.1177/1555343411415795

Department of Defense (DoD). (2012). *Test and evaluation management guide*. Retrieved from https://acc.dau.mil/docs/temg/Test%20and%20Evaluation%20Management%20Guide,%20December%202012,%206th%20Edition%20-v1.pdf

Henningsson, K. & Wohlin, C. (2004). Assuring fault classification agreement – an empirical evaluation. In *Proceedings of the 2004 International Symposium on Empirical Software Engineering*. doi: 10.1109/ISESE.2004.1334897

Holness, K. S. (2016). Content analysis in systems engineering acquisition activities. In *Proceedings of the 13th Annual Acquisition Research Symposium* (*Vol. 1)* (pp. 57–62). Monterey, CA: Naval Postgraduate School. Retrieved from http://www.acquisitionresearch.net/files/FY2016/SYM-AM-16-025.pdf

International Council on Systems Engineering (INCOSE). (2015). *Systems engineering handbook: A guide for system life cycle processes and activities*. Hoboken, NJ: John Wiley and Sons.

Jackson, P. (2009). *Getting design right: A systems approach* [Online version]. Retrieved from https://doi.org/10.1201/b15802-8

Kamongi, P., Kotikela, S., Gomathisankaran, M., & Krishna, K. (2013). A methodology for ranking cloud system vulnerabilities. In *Proceedings of the Fourth International Conference on Computing, Communications and Networking Technologies (ICCCNT)* (pp. 1–6). doi: 10.1109/ICCCNT.2013.6726854

Kenett, R. S., & Baker, E. R. (2010). *Process improvement and CMMI for systems and software* [Online version]. Retrieved from https://doi.org/10.1201/9781420060515-c6

Kossiakoff, A., Sweet, W. N., Seymour, S. J., & Biemer, S. M. (2011). *Systems engineering principles and practice*. Hoboken, NJ: John Wiley and Sons.

Krippendorff, K. (2013). *Content analysis: An introduction to its methodology*. Thousand Oaks, CA: Sage.

Lafond, D., Vallieres, B. R., Vachon, F., St Louis, M., & Tremblay, S. (2015). Capturing nonlinear judgment policies using decision tree models of classification behavior. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *59*(1), 831–835. doi: 10.1177/1541931215591251

Lehto, M. R., Nah, F. H., & Yi, J. S. (2012). Decision-making models, decision support, and problem solving. In G. Salvendy (Ed.), *Handbook of human factors and ergonomics* (pp. 192–242). Retrieved from http://onlinelibrary.wiley.com/book/10.1002/9781118131350

Leszak, M., Perry, D. E., & Stoll, D. (2002). Classification and evaluation of defects in a project retrospective. *Journal of Systems and Software, 61*(3), 173–187. doi: https://doi.org/10.1016/S0164-1212(01)00146-7

Linkov, I., Satterstrom, F. K., & Fenton, G. P. (2009). Prioritization of capability gaps for joint small arms program using multi-criteria decision analysis. *Journal of Multi-criteria Decision Analysis*, *16*, 179–185. doi: 10.1002/mcda.446

Memorandum of agreement on multi-service operational test and evaluation (MOT&E) and operational suitability terminology and definitions. (2010). Retrieved from http://www.public.navy.mil/cotf/OTD/OTA%20MOTE%20MOA.pdf

Miles, M. B., & Huberman, A. M. (1994). *Qualitative data analysis*. Thousand Oaks, CA: Sage.

Naval Air Warfare Center Training Systems Division. (n.d.). Deficiency reporting for training system testing. Retrieved from http://www.navair.navy.mil/nawctsd/Resources/Library/Acqguide/testingdeficiencyreporting.htm

Patton, M. Q. (2015). *Qualitative research & evaluation methods*. Thousand Oaks, CA: Sage.

Tao, Z., & Simons, M. (2012). A methodology for addressing operational shortfalls associated with terminal operations. In *Proceedings of the 2012 Integrated Communications, Navigation and Surveillance Conference* (pp. I31–I38). doi: 10.1109/ICNSurv.2012.6218413

Wasson, C. S. (2006). *System analysis, design, and development*. Hoboken, NJ: John Wiley and Sons.

Wickens, C., & Carswell. C. M. (2012). Information processing. In G. Salvendy (Ed.), *Handbook of human factors and ergonomics* (pp. 117–161). Retrieved from http://onlinelibrary.wiley.com/book/10.1002/9781118131350

Appendix A: Experiment Deficiency List

Issue	Title	Issue Priority Assigned by Test Personnel	Assigned To	Issue
1	Coldstart media missing	III	Prime Contractor	Government and contractors agreed to provide software coldstart media as part of technical data package during this test event. The coldstart media were not available. Coldstart media contain all of the software installation files and instructions. These are needed to restore a computer to full operational capability, in the event that the computer hard drive needs to be replaced.
2	Can't play back recorded mission	III	Prime Contractor	During final government testing, flew the designated segment per the test procedures and recorded the flight successfully. When I tried to play back the recording, the computer crashed. After system reboot, tried playback again but it didn't play smoothly. John said that during contractor testing, his recorded segment didn't play back smoothly. But, after updating the software code, it played back with no problem.
3	Instructor Station page for Nav Dis mismatch	III	Prime Contractor	The Instructor Station page for the navigation displays doesn't match the panels in the simulator cockpit.
4	Lighting system mismatch	III	Prime Contractor	The following button panels in the cockpit are fully bright when in "Night" flying mode: *Door panel *Flight Control Unit panel They should be dim, since at full bright in the darkened cockpit, they give off a lot of light. The door panel has a status button for every door on the aircraft to indicate if the door is open or closed. The flight control unit panel has buttons, knobs and displays for adjusting different aspect of aircraft attitude (pitch, roll, bank angles, etc.) and auto-pilot controls.
5	Missing video channels on surveillance	III	Prime Contractor	Only one of the three video channels for the surveillance PC came up and they were out of focus. The surveillance PC is connected to, and receives inputs from, three cameras that are installed in the simulator—one for cockpit views and two for out the window views. These duplicate cameras that are installed in real aircraft.
6	Missing Battery Indicator	II	Prime Contractor	The battery indicator is missing from the battery panel. It was ordered, just hasn't arrived yet. Will be installed when part is received.
7	Headset Mic Problem	II	Prime Contractor	Testers brought their own headsets with them. These headsets are ones they used in actual aircraft that this training device is designed to simulate. When plugged in, the person sitting in the pilot seat could hear the person at the instructor station, but not hear the person in the co-pilot seat. Tried various microphone adjustments, and different comms channels, but problem still there.

Issue	Title	Issue Priority Assigned by Test Personnel	Assigned To	Issue
8	Instructor Station–Screen capture software test incomplete	II	Prime Contractor	Can't complete the Windows Snipping Tool configuration test until printer is installed and selected within the software. Printouts of screen captures are used during debrief sessions between instructors and student pilots and co-pilots.
9	Digital Map malfunction	II	Prime Contractor	When I activated the "Map Fail" malfunction at the Instructor Station, the MAP1 and MAP2 failure advisories did not display on the control display units in the cockpit.
10	Ice Shedding/Removal	III	Prime Contractor	During a run of the basic flight scenario in winter conditions, the expected time for the ice to shed and be removed from the wings is 9 minutes, per the test procedure. Actual observed time during the test was 7 minutes, which is off by 2 minutes.
11	Gross Weight	III	Prime Contractor	Gross weight of the aircraft shown on the Instructor Station was 300 lb more than what was shown on the cockpit's control display units.
12	Engine Fire Extinguisher malfunction buttons	III	Prime Contractor	After using the left and right engine fire extinguisher malfunction buttons, pressing them again should remove the malfunctions and the button should go clear. Instead it stayed red and the malfunction did not clear.
13	Flap display not working	II	Prime Contractor	After moving the flap lever, the flap display did not indicate the change in flap position.
14	Visual Scene–Distorted Golden Gate Bridge	III	Subcontractor 2	On approach to Golden Gate Bridge, the Golden Gate Bridge seemed to be shimmering. Suspect the model in the database is the problem.
15	Visual Scene–Time of Day mismatch	II	Subcontractor 2	The time of day in the visual system doesn't match the system clock. For example, if the time of day on the system clock is morning, the visual will look like nighttime.
16	Missing panel gap covers	III	Prime Contractor	On the panels next to both the pilot and co-pilot seats, there are places where there no buttons, knobs or switches. The metal covers for these panel gaps have not yet been installed. Since these covers are non-functioning components, their installation was not required for trainer testing. However, they need to be installed prior to the government acceptance decision.
17	Incorrect weather depiction	II	Prime Contractor	From the Instructor Station, I placed a light rainstorm into the visual scene. However, on weather radar display in the cockpit, this storm was colored yellow, not green.
18	Cross winds setup	II	Prime Contractor	Test 2.8.2 Steps 14 & 15: Could not setup cross winds on the Instructor Station, so these two steps were not completed. Needs to be retested after the problem is fixed.
19	Night FLIR not working	II	Prime Contractor	Test 4.3: During testing in night conditions, the heat source signatures for anything moving did not show up on the FLIR.

Issue	Title	Issue Priority Assigned by Test Personnel	Assigned To	Issue
20	Visual scene–Ownship shadow	III	Subcontractor 2	While flying over water in the day, Ownship's shadow over the water was not aligned with aircraft position and sun position in the visual scene.
21	Visual database–Selectable storms on Instructor Station	III	Subcontractor 2	There are only eight selectable storms currently on the Instructor Station. The requirement is for a minimum of ten storms.
22	Trainer automatic power shutdown did not work	II	Prime Contractor	When power from the main circuit breaker of the training room is removed, the trainer's Automatic Power Shutdown procedure did not work. After power was disconnected at the breaker and the uninterruptable power supply (UPS) kicks in, the computers should have run through an automatic shutdown procedure, shutting them down one by one. However, this didn't happen, and the trainer stayed fully powered.
23	No audio captured in mission recording	III	Prime Contractor	No audio transmissions were recorded for mission playback. Recorded 5 minutes of a mission with radio transmissions made by the pilot.
24	Visual–Aircraft searchlight not seen at night	III	Subcontractor 2	When the night visual database was used for the aircraft sitting on the runway at Los Angeles airport, the tester set the cockpit for night mode, then pressed the button for the pilot side search light. This light could be seen in the visual scene out the window (OTW). However, when the co-pilot side search light button was pressed, the pilot search light disappeared from the visual scene, when it should have stayed on, and the co-pilot search light could not be seen OTW.
25	Weather visual scene and cockpit display mismatch	II	Prime Contractor	From the Instructor Station, I placed a light rainstorm into the visual scene. However, on weather radar display in the cockpit, this storm was colored yellow, not green.

THIS PAGE INTENTIONALLY LEFT BLANK

Appendix B: Demographics Questionnaire

Please provide us with some information about you. This information is requested so that we may document the qualifications of our study participants. If an item does not apply to you, indicate that it is not applicable by marking NA. **There are four questions with subparts. Please answer all questions**.

1. Your NPS Degree Program Curriculum Number and Specialization Track: (Example: 580 Combat Systems)

2. Your Service, Rank, Most Recent Job Title before coming to NPS: (Example: U.S. Navy, LT/O3, EDO. Example: Army—Singapore, Civilian, Program Manager)

3. Have you had past experience evaluating system T&E data? Please enter Yes or No. If No, skip to question 4.

3a. If yes, approximately how many months or years, in total, did you spend evaluating system T&E data?

3b. For which system(s) did you evaluate T&E data?

4. Educational background:

4a. List your current bachelor's degree(s): (Example: BA American Studies)

4b. List your current master's degree(s): (Example: MS Industrial Engineering)

4c. List your current doctoral degree(s) (PhD, MD, JD, EdD, etc.): (Example: PhD Industrial Engineering)

THIS PAGE INTENTIONALLY LEFT BLANK

Appendix C: Classification Strategies Questionnaire

Please provide us with information about the steps your took to create your final set of categories and priority order for the deficiencies.

There are eight questions. Please answer all questions.

1. Describe the similarities and/or differences you noticed between the deficiencies.

2. Describe anything else you considered when creating the categories.

3. List the questions you tried to answer when you assigned each deficiency to one or more categories. Please be specific on the actual things you did and decisions you made.

4. Describe other things you considered as you assigned **each deficiency to one or more categories**. Please be specific on the actual things you did and decisions you made.

5. If there was more than one category that a deficiency could belong to, describe how you made your final decision for that deficiency's category. Please be specific on the actual things you did and decisions you made.

6. Describe how you assigned **a priority number each deficiency**. Please be specific on the actual things you did and decisions you made.

7. If there was more than **one priority ranking** that a deficiency could belong to, describe how you made your final decision for that deficiency's category. Please be specific on the actual things you did and decisions you made.

8. Describe anything from your previous training or T&E experience you used when creating the **categories**.

9. Describe anything from your previous training or T&E experience you used when creating the **priority order of deficiencies**.

THIS PAGE INTENTIONALLY LEFT BLANK

Appendix D: Workload Assessment Questionnaire

Questionnaire 3: Workload Assessment

Please rate the different workload demands you experienced during the task of categorizing and prioritizing the deficiencies. Use the following scale from 1 to 10, as defined in the tables below. Type your choice for each demand under **"Enter Data Here"**

	LOW / POOR									HIGH / GOOD
Mental Demand	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*
Temporal Demand	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*
Performance	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*
Effort	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*
Frustration Level	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*

NASA TLX Work Load Rating-Scale Descriptions:

Title	Descriptions	Enter Data Here
Mental Demand	How much mental and perceptual activity was required (e.g., thinking, deciding, remembering, looking, searching, etc.)? Was the task easy or demanding, simple or complex, exacting or forgiving?	
Temporal Demand	How much time pressure did you feel due to the time constraints to categorize and prioritize the deficiencies? Was the pace slow and leisurely or rapid and frantic?	
Performance	How successful do you think you were in accomplishing the goals of the task? How satisfied were you with your performance in accomplishing the goals?	
Effort	How hard did you work (mentally) to accomplish your level of performance?	
Frustration Level	How insecure, discouraged, irritated, stressed, and annoyed versus secure, gratified, content, relaxed and complacent did you feel during this tas	

THIS PAGE INTENTIONALLY LEFT BLANK

Appendix E: Perceived Value Questionnaire

Questionnaire 5: Perceived Value

Please rate the value of categorizing the deficiencies before prioritizing. Use the following scale from 1 to 10, as defined in the tables below.

	LOW									HIGH
Value of Categories	1	2	3	4	5	6	7	8	9	10
Impact of Categories	1	2	3	4	5	6	7	8	9	10

Perceived Value Rating-Scale Descriptions:

Title	Descriptions	Enter Data Here
Value of Categories	In general, when evaluating T&E deficiencies for any system, how valuable do you think it is to first categorize the deficiencies before prioritizing them?	
Impact of Categories	How much impact did the categories have on the priority order you came up with?	

Enter any additional comments here:

THIS PAGE INTENTIONALLY LEFT BLANK

Appendix F: Complete Categorization Results

Issue #	Issue Title	Priority Assigned by Test Personnel	Category Subject 1 (T)	Category Subject 3 (T)	Category Subject 5 (T)	Category Subject 2 (NT)	Category Subject 4 (NT)
6	Missing Battery Indicator	II	Hardware	Part II. Hardware functionality missing. Testing not completed.	Ancillary, Priority D	Physical component, Part III	Minor
7	Headset Mic Problem	II	Hardware	Part II. Additional information required on availability of workaround and what the contract specified. Potential to be a Part I.	Ancillary, Priority D	Interface, Part II	Major
8	Instructor Station–Screen capture software test incomplete	II	Hardware	Part II. Hardware functionality missing. Testing not completed.	Instructor, Priority B	Data capture, Part III	Minor
9	Digital Map malfunction	II	Simulation Software	Part II. Software bug.	Cockpit, Priority B	Procedure mismatch, Part II	Critical
13	Flap display not working	II	Hardware	Part II. Software bug.	Cockpit, Priority B	Procedure mismatch, Part I	Minor
15	Visual Scene–Time of Day mismatch	II	Simulation Software	Part II. Software bug.	Visual, Priority C	Visual system delta, Part III	Critical
17	Incorrect weather depiction	II	Simulation Software	Part II. Software bug.	Cockpit, Priority C	Visual system delta, Part III	Critical
18	Cross winds setup	II	Simulation Software	Part II. Software bug.	Instructor, Priority A	Procedure mismatch, Part I	Minor
19	Night FLIR not working	II	Simulation Software	Part II. Software bug.	Visual, Priority B	Physical component,	Minor

Issue #	Issue Title	Priority Assigned by Test Personnel	Category Subject 1 (T)	Category Subject 3 (T)	Category Subject 5 (T)	Category Subject 2 (NT)	Category Subject 4 (NT)
						Part I	
22	Trainer automatic power shutdown did not work	II	Hardware	Part II. Software bug? Functionality missing. Testing not completed.	Ancillary, (safety), Priority A	Physical component, Part I*	Major
25	Weather visual scene and cockpit display mismatch	II	Simulation Software	Part II. Software bug.	Cockpit, Priority C	Visual system delta, Part III	Minor
1	Coldstart media missing	III	Technical Software	Part III. Hardware functionality missing. Testing not completed.	Data, priority A	Physical component, Part I	Critical
2	Can't play back recorded mission	III	Technical Software	Part III. Software bug.	Instructor, Priority A	Data capture, Part I	Major
3	Instructor Station page for Nav Dis mismatch	III	Hardware	Part III. Software bug.	Instructor, Priority A	Procedure mismatch, Part III	Minor
4	Lighting system mismatch	III	Hardware	Part III. Hardware functionality missing. Testing not completed.	Cockpit, Priority D	Physical component, Part II	Minor
5	Missing video channels on surveillance	III	Hardware	Part III. Hardware functionality missing. Testing not completed.	Visual, Priority C	Interface, Part III	Major
10	Ice Shedding/ Removal	III	Simulation Software	Part III. Software bug.	Visual, Priority C	Procedure mismatch, Part III	Major

Issue #	Issue Title	Priority Assigned by Test Personnel	Category Subject 1 (T)	Category Subject 3 (T)	Category Subject 5 (T)	Category Subject 2 (NT)	Category Subject 4 (NT)
11	Gross Weight	III	Simulation Software	Part III. Software bug.	Instructor, Priority B	Procedure mismatch, Part III	Critical
12	Engine Fire Extinguisher malfunction buttons	III	Hardware	Part III. Software bug.	Cockpit, Priority A	Procedure mismatch, Part I	Safety/ critical
14	Visual Scene– Distorted Golden Gate Bridge	III	Simulation Software	Part III. Software bug.	Data priority C	Visual system delta, Part III	Safety/ critical
16	Missing panel gap covers	III	Hardware	Part III. Non-functional hardware deficiency.	Ancillary, Priority C	Physical component, Part II	Minor
20	Visual scene– Ownship shadow	III	Simulation Software	Part III. Software bug.	Ancillary (OTW), Priority C	Visual system delta, Part III	Minor
21	Visual database– Selectable storms on Instructor Station	III	Simulation Software	Part III. Software bug.	Instructor, Priority D	Visual system delta, Part III	Major
23	No audio captured in mission recording	III	Technical Software	Part III. Additional information required on availability of workaround and what the contract specified. Potential to be a Part I.	Instructor, Priority A	Data capture, Part I	Critical
24	Visual– Aircraft searchlight not seen at night	III	Simulation Software	Part III. Software bug.	Visual, Priority B	Visual system delta, Part III	Major

THIS PAGE INTENTIONALLY LEFT BLANK

Appendix G: Part III Prioritization Results

Issue #	Issue Title	Priority Assigned by Test Personnel	Priority for Subject 1 (T)	Priority for Subject 3 (T)	Priority for Subject 5 (T)	Priority for Subject 2 (NT)	Priority for Subject 4 (NT)
1	Coldstart media missing	III	3	15	2	1	3
2	Can't play back recorded mission	III	1	16	6	1	10
3	Instructor Station page for Nav Dis mismatch	III	9	23	3	3	18
4	Lighting system mismatch	III	7	13	25	2	23
5	Missing video channels on surveillance	III	8	14	15	3	14
10	Ice Shedding/Removal	III	7	22	16	3	11
11	Gross Weight	III	6	18	13	3	4
12	Engine Fire Extinguisher malfunction buttons	III	3	17	7	1	2
14	Visual Scene– Distorted Golden Gate Bridge	III	12	21	21	3	1
16	Missing panel gap covers	III	10	25	17	3	25
20	Visual scene– Ownship shadow	III	8	20	20	3	20
21	Visual database– Selectable storms on Instructor Station	III	10	24	23	3	12
23	No audio captured in mission recording	III	2	12	5	1	7
24	Visual–Aircraft searchlight not seen at night	III	9	19	14	3	13

ACQUISITION RESEARCH PROGRAM
GRADUATE SCHOOL OF BUSINESS & PUBLIC POLICY
NAVAL POSTGRADUATE SCHOOL
555 DYER ROAD, INGERSOLL HALL
MONTEREY, CA 93943

www.acquisitionresearch.net