SYM-AM-18-044

PROCEEDINGS
OF THE
FIFTEENTH ANNUAL ACQUISITION RESEARCH SYMPOSIUM

WEDNESDAY SESSIONS
VOLUME I

Acquisition Research:
Creating Synergy for Informed Change

May 9–10, 2018

Published April 30, 2018

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

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

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

Karen Holness—is an Assistant Professor in the Department of Systems Engineering at the Naval Postgraduate School (NPS) in Monterey, CA. She holds a BS, MS, and PhD in Industrial Engineering from the University at Buffalo. Prior to NPS, she worked as a Navy civilian in the acquisition workforce for eight and a half years in various industrial engineering, systems engineering and human systems integration roles. She also previously worked for three years as an Industrial Engineer at Corning, Incorporated in Corning, NY. [kholness@nps.edu]

Rabia H. Khan—is a Faculty Research Associate in the Systems Engineering Department at the Naval Postgraduate School (NPS) in Monterey, CA. Khan's research background includes studies in systems engineering competency modeling and development, cognitive processing, content analysis, energy studies, curriculum development and measuring self-efficacy within the field of systems engineering. She earned her BS in Psychobiology from the University of California, Davis and an MS in Engineering Systems from NPS. [rhkhan@nps.edu]

Gary Parker—is a Faculty Research Associate in the Systems Engineering Department at the Naval Postgraduate School (NPS) in Monterey, CA, with 41 years of combined military, industry, and government experience in intelligence, systems engineering, and defense systems acquisition. His research areas include systems behavioral modeling, complex adaptive systems, and system of systems engineering. Parker's education includes a BS in Aerospace Engineering from the University of Colorado, an MS in Business Administration from Boston University, and MS degrees in Systems Engineering and Physics from NPS. He is currently a PhD student in the Systems Engineering Department. [gwparker@nps.edu]

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.

Introduction

Like other data analysis efforts within a typical Department of Defense (DoD) acquisition program, test and evaluation (T&E) data analysis efforts are impacted by constraints of program cost, schedule, and resource availability. 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. The analysis of T&E data is further impacted by the inherent mental models, heuristics, and biases each government and contractor engineer brings to working on the same dataset based on their individual backgrounds and experience.

Holness (2016) described the potential for 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. This current empirical study 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 paper, variations on the terms *deficiency*, *discrepancy*, *anomaly*, *issue*, *problem*, *failure*, and *fault* are considered synonymous and are used interchangably.

Literature Review

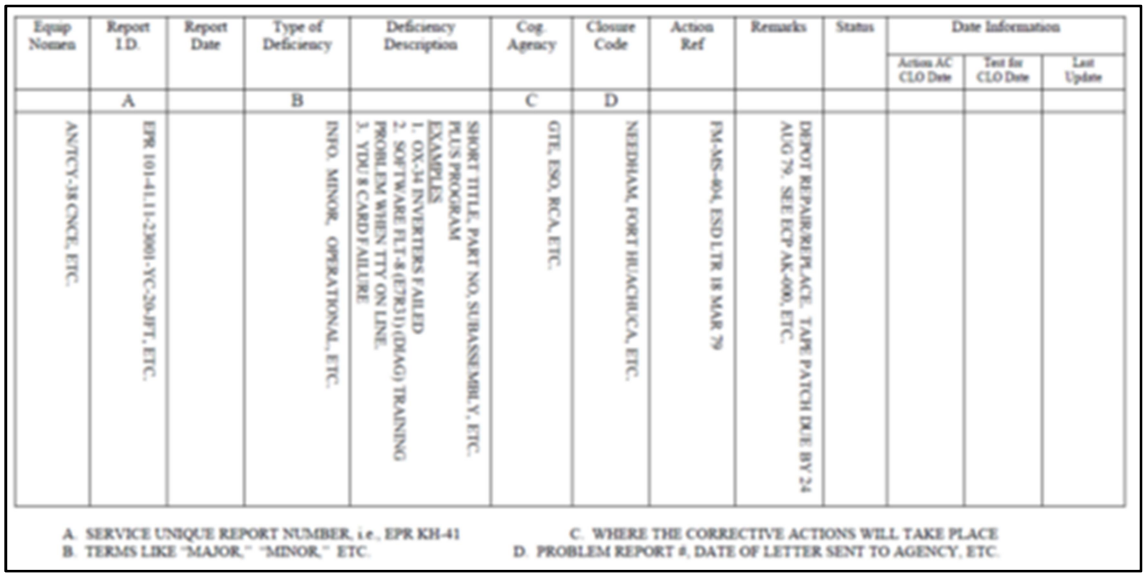
The standard process for conducting a T&E event involves adherence to a pre-established T&E plan that supports either system verification or validation activities with an approved set of test procedures. After executing the test procedures and recording the results, observed anomalies are analyzed and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

Kossiakoff et al. (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 deficiency 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 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 systems engineering team members also evaluate the qualitative and quantitative data contained in written text and the deficiency codes in discrepancy descriptions 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						
ANTCY-38 GNCE, ETC.	EPR 101-41.111-23001-YC-20-IFT, ETC.		INFO, MINOR, OPERATIONAL, ETC.	SHORT TITLE, PART NO, SUBASSEMBLY, ETC. PLUS PROGRAM EXAMPLES 1. OX-34 INVERTERS FAILED 2. SOFTWARE FLT-8 (E783) (DIAG) TRAINING PROBLEM WHEN TTY ON LINE 3. YDU 8 CARD FAILURE	GTE, ESD, 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-600, 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
(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 (2017) website, "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) *Test and Evaluation 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. 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.

There is variability in how best to tailor an approach for a specific work domain. The common approach across a variety of deficiency classification and prioritization tasks is some combination of calculated numerical scores and human judgment. Of particular interest in this research is the creation and use of additional classification categories to complete a prioritization task. This emphasis on embedding classification within prioritization

warrants a discussion about the fundamentals of content analysis as a categorization process for qualitative data.

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, and 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 reliability, 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 objective of this research study is to investigate different ways that system issues with assigned deficiency classifications are prioritized for resolution. Of particular interest are the strategies individuals use to prioritize a list of deficiencies for resolution, with or without prior knowledge of the content analysis methodology. The next section describes the design of this study.

Methodology

All research design and execution activities were completed at the Naval Postgraduate School (NPS) by the authors of this report. 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, and
- designed to target a specific deficiency prioritization solution.

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. Additional students were recruited from the following curricula: Naval/Mechanical Engineering (Total Ship Systems Engineering track), Systems Acquisition Management, and 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 both conditions, 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 resident systems engineering 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 et al. (2015), 60 university students performed a radar contact classification task in a naval air-defense scenario using a simulated combat control system microworld. Finally, in the Cropp, Banks, and Elghali (2011) study, 30 industry professionals reviewed hypothetical case studies and rated potential risks associated with each one.

Data Collection Method

A pilot study was conducted prior to the main experiment. 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, some classroom visits 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 while executing a set of approved simulator test procedures. The deficiency list provided in the T&E report contained 25 issues. A brief description was provided for each issue, 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. The T&E report provided definitions of all of the NAWCTSD classifications for each subject's reference.

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; they were asked to assume that one would be created for each deficiency after the priority order for resolution was completed. The subjects were asked to complete the categorization and prioritization task within one hour using the provided T&E report as a reference and working with the list of deficiencies in a Microsoft Excel spreadsheet. A pen and paper were 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.

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.

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.

Results Summary

The participants included one NPS employee and four NPS students. Two students were from the SE curriculum, and two were from the Systems 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.

Categorization Results

Table 1 shows a sample of the results of the categorization exercise for the flight simulator Part II issues. The results were grouped by training condition 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 Problems	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

It is noteworthy 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.

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 assigning issues to the created categories. Given the aforementioned observations on the categorization strategies used by the test subjects, it appears that subjects used heuristics to focus on high-level attributes of the system, perhaps as a way to manage and consolidate the data in a meaningful way. Each subject made a judgment of circumstance, scope, and criticality using the provided descriptions of each issue and their own interpretations and mental model of each issue. Despite the similarities in some of the category names, each person's working definition of these categories was different enough to preclude the same issues all being assigned to the same categories. It is difficult to tell what their categories would have looked like if they had been specifically instructed to use the test personnel prioritizations. Based on these results, the impact of the content analysis training was inconclusive.

Prioritization Results

Each subject was asked to first categorize the issues and then prioritize the issues for remediation. Table 3 lists the assigned priority numbers for the Part II issues. The results were grouped by subject training condition 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 10 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. It was expected that all of the Part II issues would appear within the top 11 rankings of the prioritization list had the subjects leveraged the priority from the test personnel. 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 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).

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 table. 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.

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.

Figure 2 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.

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

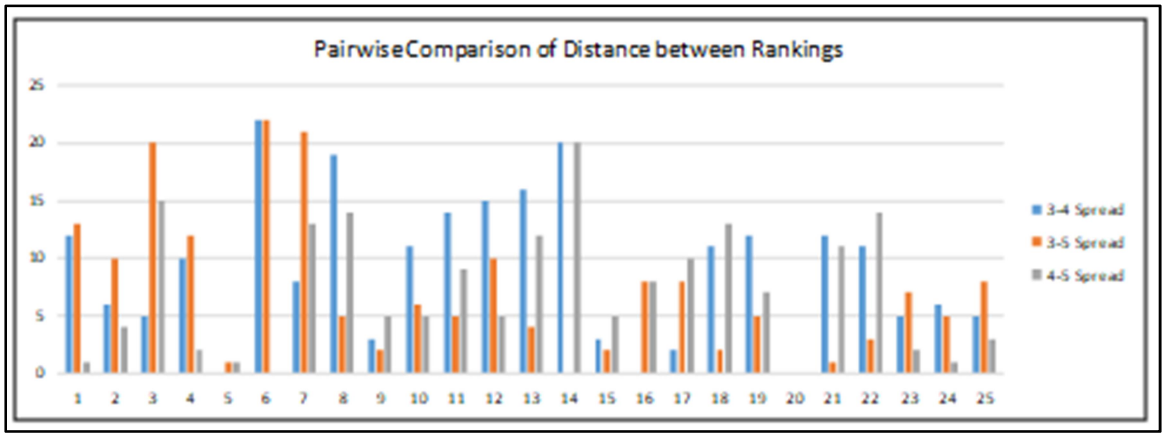


Figure 2. Pairwise Comparison of Distance Between Rankings by Issue

Given these findings, no 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.

Classification Strategies Questionnaire Results

In the classification questionnaire, subjects were asked to describe the rationale they used to create categories and assign a resolution priority number to each issue. Figure 3 shows the reported answers summarized into seven categories.

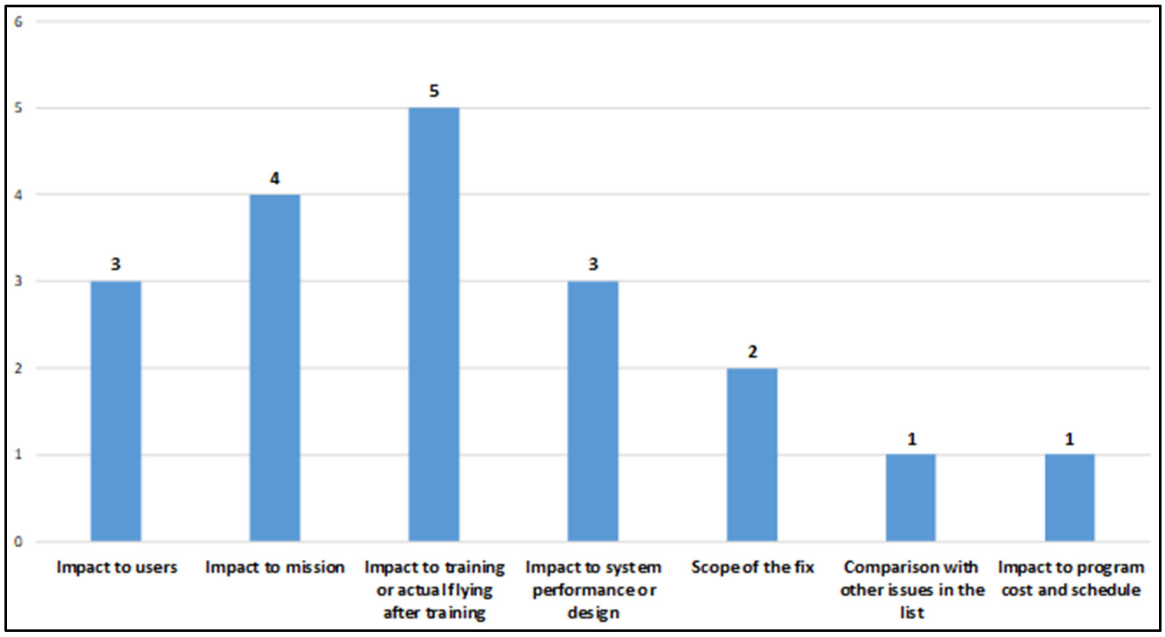


Figure 3. Counts of Reported Rationale

Impact to users, impact to mission, and impact to training or actual flying after training seem similar and could possibly 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 subject was in the non-training condition but did have a background in the T&E domain.

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

Workload Assessment Questionnaire Results

Table 5 summarizes the subject responses to the workload assessment questionnaires. Subjects were asked to rate their perceived level of workload on a number of factors, on a scale from 1 to 10, with "1" reflecting a "poor" level and "10" being a "good" level.

Table 5. 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, Subjects 2 and 3 exceeded the allotted hour by 8 minutes and 14 minutes, respectively. Subjects 1 and 4, who reported the highest mental demand and temporal demand scores, 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.

Perceived Value Questionnaire Results

The subjects rated 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" reflecting a "low" perceived value and "10" being a "high" perceived value. Table 6 summarizes these responses. No significant differences were observed between those in the training condition and the control condition on the value scores.

Table 6. 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. Subject 2 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.

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 they did not use the issue priority assigned by the test personnel. In this subject'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 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 results that are more consistent across subjects. It would be interesting to revise this research study to investigate how subjects assign issues to pre-defined technology-based and priority-based categories provided to them. Another variation would be to pre-assign issues to such categories and 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.

References

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

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

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.

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

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 (MOA) 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 (NAWCTSD). (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.

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

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

www.acquisitionresearch.net