Proceedings of the Annual General Donald R. Keith Memorial Conference
West Point, New York, USA
May 2, 2019
A Regional Conference of the Society for Industrial and Systems Engineering

Codifying Best Practices of GUI Testing Within A Standardized Framework

Marina Camacho, Sean Devine, Austin Harry, Quinten Parker, Gabrielle Purnell, and Jeffrey Demarest

Department of Systems Engineering
United States Military Academy, West Point, NY

Corresponding author: gabrielle.purnell@westpoint.edu

Author Note: Cadets Camacho, Devine, Harry, Parker and Purnell are working for the National Assessment Group (NAG) as a part of the USMA Capstone program with the Department of Systems Engineering and Operations Research Center. Mr. Jeffrey Demarest is the Capstone Advisor to the Capstone team.

Abstract: Graphical User Interfaces (GUIs) must be evaluated based on their usability and suitability in order to maximize the relationship between the human and the GUI. This research seeks to develop a standardized testing process that enables analysts to test and score the relationship. The research draws upon several fields including human-computer interaction, human-factors engineering, usability and suitability as a means of capturing GUI testing best practices. More specifically, testing the aspect of Human-in-the-Loop (HITL) in systems acts as the primary means of interaction between the background functionality of the user and computer operated system. The study of GUI testing involves a direct applicability to the professional careers of Army Officers as it fosters an understanding of the effectiveness of equipment for soldiers on the battlefield. This research applies the principles of GUI testing to the realm of defense-based systems to produce a standard GUI testing framework for use within the defense testing community.

Keywords: Graphical User Interface, Usability, Suitability, Testing, Human-in-the-Loop, Human-Machine Interaction

1. Introduction

Graphical User Interfaces (GUIs) are nothing more than "a computer program designed to allow a computer user to interact easily with the computer, typically by making choices from menus or groups of icons" (Merriam-Webster, 2018). They are used every day, whether it be on cell phones, tablets, computers, smart watches, or vehicle dashboards. GUIs are so ubiquitous that most people do not realize the technical name for them.

The overall effort of this project is to develop a standardized testing process that enables analysts to test and score the interaction between humans and GUIs. The well-developed framework that the team will design, create, and implement has the opportunity to impact soldiers in the field and in combat, as it will allow analysts to determine the most functional and relative GUIs that satisfy the military's wants and needs. This research develops a testing framework that will be useful for every soldier and every platform that the military has to offer. Whether that be global positioning systems (GPS), handheld support systems, or aircraft dashboards, the testing framework accounts for all GUIs.

1.1 Background

The National Assessment Group's (NAG) current approach to analyzing the Human-in-the-Loop (HITL) aspect of GUIs consists of conducting several user surveys to determine the usability between the human and the GUI. Due to the NAG's lack of measures focused on user interface assessment, their data analysis of the HITL aspect is very limited. The NAG's standard approach for collecting user feedback includes the use of the System Usability Scale (SUS). Similar to the well-known Likert Scale, the SUS "is a simple, ten-item scale giving a global view of subjective assessments of usability" (Usability Evaluation in Industry, p. 191). The usability scale allows the user to select from a range of 'Strongly Agree' to 'Strongly Disagree' based on their performance on a specific task. While the NAG's procedure is simple and not time intensive, they recognized this approach may result in fielding of a system that is not optimal for the HITL interaction.

As part of the current NAG approach for assessing GUIs, they conduct field testing that includes small sample sizes. The disadvantage of having a small sample size when collecting data is that the data may not fully represent the product, compromising the NAG's test accuracy. Another limitation of the NAG's current strategy for assessing GUIs is only focusing on the usability aspects between the human and the GUI. A narrow view in the analysis approach creates more limitations for creating a standardized, generalizable solution when analyzing a GUI. This research identifies a useful way to incorporate the aspects of testing suitability and usability when analyzing the interaction between the human and the GUI.

1.2 Problem Statement

The goal of this research is to develop a standardized testing process that is platform agnostic and generalizable across multiple test cases to allow analysts to test and score the interaction between the human and the GUI. The primary beneficiary of the process are NAG analysts. Analysts are the individuals that are tasked with determining the operational utility of a system. They analyze the effectiveness, suitability, and survivability of the system that is being tested.

2. Methodology

The methodology to pursue the wants and needs of the NAG followed a standard process in which general assumptions, constraints, and limitations were identified prior to the data collection. These factors drove the research and data collection in providing a clearer scope while researching and creating the literature reviews. The literature reviews were used to draw out key information in the human and GUI interaction which was used in model development.

2.1 General Assumptions, Constraints, and Limitations

In creating a framework that incorporates all the requirements that the NAG expects, this research made several assumptions. First, the team assumed the operating system functions properly when initially presented to the user. Additionally, the team assumed that there is variation in the type of users. Specifically, according to the NAG, the user could be anyone from a member of a Special Missions Unit to a cyber protection team member which makes developing a generalizable framework difficult. Finally, the team is operating under the assumption that a GUI being tested by the proposed test is assumed to be operationally functional. This research is primarily limited by the knowledge of how to use the selected software in order to sufficiently demonstrate the functions of the software to the NAG. The main functions we intend to test are the compatibility, the physical dimensions, environmental performance, effectiveness, user satisfaction, workload, and efficiency of the system.

2.2 Data Collection

Over the past few months, the team has collected data by conducting multiple stakeholder semi-structured interviews and literature reviews. The stakeholder analysis sampled civilian and military professionals who have experience in the GUI testing realm. Questions given to the interviewees addressed the policies they currently use, their techniques to evaluate interfaces, current faults within their testing process, and the different testing environments they utilize. The team found the common trend between the stakeholders was testing the effectiveness of the system, user satisfaction, and using a form of a Likert or point scale in order to score the system. These factors attribute to a holistic picture of the tested system given the relationships discussed in Figure 1. Similarly, the Likert-type scale simplifies the mathematics in scoring the given system, allowing for a uniform process across multiple tests of a given system.

The most influential stakeholder that greatly contributed to the research was a usability testing expert for Project Manager Mission Command at Aberdeen Proving Ground in Maryland. Based on her extensive research, she recommends using a four-point scale to rate specific tasks to force testers to make a decision on whether their experience was favorable or not, utilizing the Modified Cooper-Harper Cognitive Workload to measure whether the system is completing most of the work with little human input, understanding the three main usability measures (effectiveness, efficiency, and satisfaction), and identifying a potential COTS software called Morae. (P. Savage-Knepshield, personal communication, October 31, 2018).

2.3 Literature Reviews

Each cadet in the group explored application and methods used to test the HITL interactions with GUIs. The team analyzed the overall concept of HITL, different standards already in practice in the DoD, and how to convert qualitative measures to quantitative measures. The combination of standards used by the Department of Defense (DOD) and Department of the Army (DA) helped this research create a generalizable standard to use across all testing cases. In addition to the standards, the general concept of HITL and converting qualitative to quantitative measures began the process of developing the framework.

A look into the literature surrounding HITL can begin with Charles Billings, who offers a set of general principles that guide Human Computer Design (HCD) projects. Although his study, *Human-Centered Aircraft Automation,* focuses on these principles as they are applied to aircraft automation, the principles encompass the major HITL design considerations. Billings directs that effective systems must emphasize the involvement of an informed human operator. Such an operator must have the capability to monitor the system and be in command. In turn, the system itself must be able to monitor the human operator as Billings also considers human fallibility. The design of the system must incorporate the need for system predictability and each element of the system must be interconnected in order to "have knowledge of the others' intent" (Billings, 1991, p. 13).

In a similar approach to Billings, Asaf Degani and Michael Heymann offer an approach to HCD as they highlight essential elements of human-machine interactions or interfaces. Their four elements are the machine's behavior, the task specification or operational goals, the user model, and the user interface. Degani and Heymann's elements include a large emphasis on the consideration for the user involvement within the system. Degani and Heymann also take note of the required relationship between the four elements. Effective systems involve a likewise effective balance between these elements (Degani and Heymann, 2002, p. 29). This concept hints back to Billing's inherent requirement for an effective balance between the involvement of the user and the system.

Within the current testing standards and practices, the DoD and DA each have several evaluation techniques and standards that are implemented into testing. The DoD has a design criteria standard, MIL-HDBK-46855, which applies the general requirements for mission success. The DoD also uses the Scientific Test and Analysis Techniques Center of Excellence (STATCE) under the stewardship of the Air Force Institute of Technology and has recently implemented the National Institute of Standards and Technology's (NIST) standard for a holistic approach for standards. The NIST uses the National Voluntary Conformity Assessment System Evaluation (NVCASE). The DA's standards include the AR-73-1, which contains the regulations for testing and evaluating in the Army. The two types of operational test and systems evaluation techniques relevant to Information Technology systems in AR-73-1 are the user acceptance test (UAT) and the supplemental site test (SST). The final DA standard is the Human Factors Engineering Data Guide for Evaluation (HEDGE), which is a twelve-step process for evaluations. All of the above testing standards were taken into consideration in the development of the GUI Testing Best Practices & Procedures.

The research conducted on the topic of converting qualitative measures to quantitative measures requires an understanding of the definitions of qualitative data and quantitative data. Qualitative data is information concerned with the understanding of human behavior from the informant's perspective (McLeod, 2017). The informant is the individual taking the given test, utilizing the system, or conducting a function. In relation to this research, this would apply to some qualities such as a rugged feeling of a GUI in the hands of a soldier testing and utilizing the system. Qualitative data "assumes a dynamic and negotiated reality" from the perspective of the individual giving his or her intuitive response to the given scenario (McLeod, 2017). Qualitative data is collected through observations and interviews where the individual reports the language in his or her own words. This implies that the individual's understanding of his or her experience may not be consistent throughout repeated trials over the course of a given population. Quantitative data is information that assumes a fixed and measurable reality (McLeod, 2017). This applies directly to the capstone in being able to measure the effectiveness and suitability of the given system and the particular GUI associated with that system. Researchers collect quantitative data through measuring things and analyzing phenomena through numerical comparisons and statistical inferences (McLeod, 2017).

2.4 Model Development

The System Scope model, shown in Figure 1, was created to serve as the basic foundation for testing framework methodology to analyze the two distinct relationships occurring: the human/user and the GUI and the GUI with the operating system. The initial approach to defining the scope characterized the GUI as strictly an interaction or connector/edge in the machine-human network. The approach characterized the GUI not as just and edge or arrow in the network in Figure 1, but as a standalone node within the dashed relationship. The dashed line in Figure 1 highlights the focus on the interaction between the human/user and the GUI itself. This is important to note as the primary focus because there exists the secondary focus of usability of the GUI and the operating system. With this approach the GUI can be seen as interacting with the operating system (OS) and the human/user separately. Based on the feedback from the client, the group defined the new scope under the assumption that a GUI being tested by the proposed test is assumed to be operationally functional. With this assumption the scope of this research is refined to focus solely on the interaction between the human/user and the GUI.

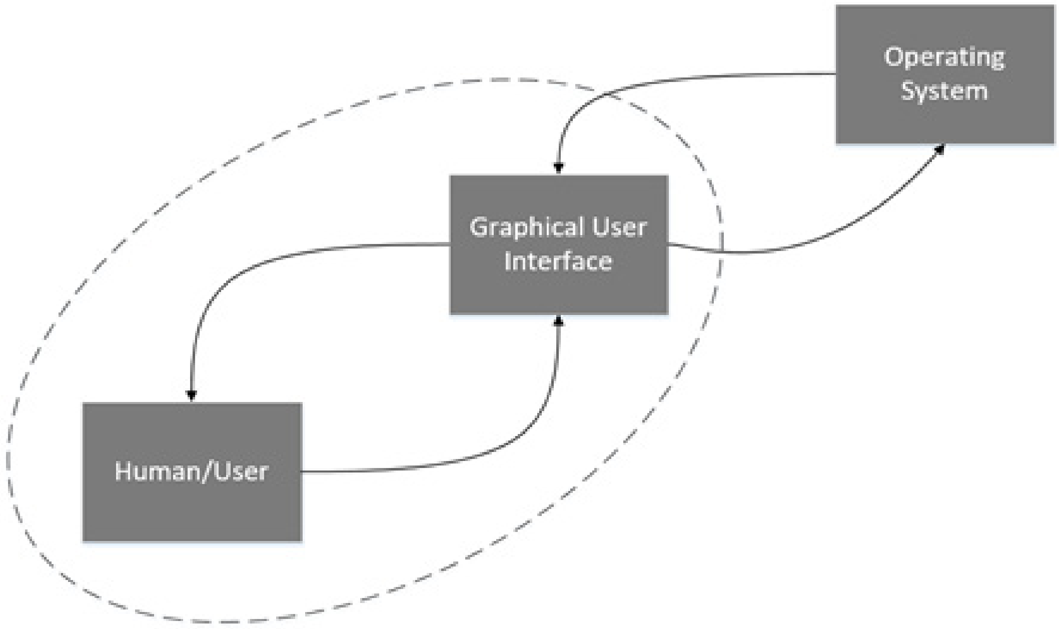

Figure 1. System Scope Model

The purpose of a functional hierarchy is to identify the functions and subfunctions of the overall system (Parnell, 2011, p. 317). The team used the functional hierarchy to help guide the concept development, design, and help identify performance measures for the efficiency and effectiveness of a GUI. To encompass the overall objective of the software, the top-level functions of the system are: 1) test the human-GUI interaction, 2) assess test data, and 3) be generalizable across multiple test cases. Within the first function, test the human-GUI interaction, the aspects of testing suitability and usability. The overall purpose of these two sub-functions is to ensure that the GUI can meet the demands of its user and its environment.

Upon completion of the functional hierarchy, the team generated a functional flow diagram. The diagram incorporated all the functions from the functional hierarchy and places them in sequential order. It created a foundation for a hard copy step-by-step process that an analyst can use to efficiently evaluate a GUI. The functional flow diagram highlights the focus on measuring both usability and suitability. The standardized testing framework allows analyst to create and conduct a test to gain an overall general understanding when evaluating and comparing GUIs.

2.5 Explanation of the Assessment Methodology

The purpose of the team's process is to teach the analyst how to test and assess a GUI by providing best practices, a step-by-step process, and method of analyzing data. Thus, the How-To Guide, in providing a standardized testing framework, will allow the analyst to employ the best practices to maximize the test and achieve high quality results. The first section of the How-To Guide will provide the analyst with the best practices for conducting a test for a GUI. These best practices are split into two main subsections: usability and suitability. Within the usability subsection, the best practices explained include task success, task time, requested assistance, and workload measurement. Furthermore, the suitability subsection explains measures of capability between the user and GUI, color blindness, physical dimensions, and the environmental performance. With best practices in mind, the How-To Guide will then continue with a step-by-step process of how to test a GUI. The final portion of the How-To Guide is the data analysis, which explains the math behind the qualitative to quantitative aspects of the testing process. The entire How-To Guide will allow the analyst to successfully evaluate and compare any type of GUI.

3. Case Study

3.1 Explanation of Case Study

The purpose of the case study is to determine the effectiveness of the team's own established testing framework in practice. The team will analyze the results to ensure that the test in the case study meets the team's functional hierarchy objectives. The team will be utilizing the GUI on an iPhone 6 Plus and accomplishing the following critical tasks: 1. Power on iPhone 2. Unlock iPhone 3. Make a call 4. Send a text message 5. Use Siri to set an appointment. The team chose this particular device given its accessibility for a test case, the ability to highlight the generalized aspects of the proposed framework, and a tailored approach to the military testing community with mobile devices in operations. These five critical tasks, although fairly simple, will allow the team to validate that the framework is capable of testing a GUI. In the case study, the team asked fifteen participants to complete the following tasks and complete the data collection sheet provided for them. The analysis sheet includes the time to complete the task, if the task was completed, if/and how many times he/she requested for assistance, the workload, five usability questions, and a suitability survey. The workload was measured using a modified Cooper-Harper cognitive workload scale from 1-10, 1 being very easy and highly desirable, and 10 being impossible (Cooper and Harper, 1969, p. 11). The suitability survey contains a set of questions regarding ergonomics, compatibility with the environment, compatibility with color blindness, power, durability, lighting and audio, and the tactical compatibility. Some example questions that assessed the suitability of the GUI were: I felt as though I can maintain positive control of this system interface at all times (where more is better), I feel as though I can quickly operate this GUI in a stress filled environment (when more is better), Color coding had a significant impact on my ability to complete the critical tasks (where less is better). These questions were scored based on a 1-5 Likert Scale (strongly disagree, disagree, neutral, agree, strongly agree). The questions ranged from a series of positive questions (where more is better) to negative questions (where less is better). The team mixed the positive and negative questions throughout so that the user did not click through each question knowing that more is always better, or vice versa. It forces the user to go through each question and analyze the best number to choose for that particular question. Once the results were recorded, the framework was analyzed and scored based on the results.

3.2 Analysis of Results

The case study proved to be very effective in the fact that it gave the team valuable feedback about the developed framework. From the case study, the team found that a positive and crucial aspect of the step-by-step framework was a well-organized flow of events. An example of this is clarifying the key tasks and the operational environment and condition before conducting any tests. This is important because the test results could be completely different considering the many different demands of various operating environments.

The team also found that within the suitability survey, the participants found confusion regarding the (+/-) system. This is where the scores differ in anticipated outcomes. Within the results the team found that participants had polar opposites in response at times that didn't follow the trend of the test. This means that if it was a question where less is better, some participants put a response of a higher number because they thought more was better. The team believes this is due to the (+/-) confusion. The team recommends having a mix of positive and negative questions throughout the survey as this serves to prevent any unintended biases from the tester and analyst regarding the GUI. A way the team can eliminate this confusion is rather than using numbers to distinguish between "strongly agree or strongly disagree," the team use the actual words "strongly agree or strongly disagree." These words can then be evaluated and changed for analysis on a later date.

4. Conclusion

The standardized testing framework encompasses a variety of the best practices of GUI testing. The analysis indicates that the most important aspects of the testing framework are identifying the GUI's critical tasks and considering the specific operating environment when evaluating the interaction between the human and the GUI. Both aspects affect the task completion time, task success, number of instances of a request for assistance, the perceived workload, and the measures of compatibility. The standardized testing framework will add value for analysts to successfully test the interaction between the human and the GUI. Without this framework, the analyst does not have a sufficient process in determining the success or failure of the GUI.

For future work, the testing process can be adjusted by using this testing framework with a more automated software. Completing a test, given a different operating environment, will provide better responses from the user when considering the usability and suitability of a GUI. The inclusion of different types of GUIs will provide better feedback and more practices when incorporated into the NAG's current testing process.

5. References

Billings, C. E. (1991). Human-centered aircraft automation: A concept and guidelines.

Cooper, G. E., & Harper Jr, R. P. (1969). *The use of pilot rating in the evaluation of aircraft handling qualities* (No. AGARD-567). ADVISORY GROUP FOR AEROSPACE RESEARCH AND DEVELOPMENT NEUILLY-SUR-SEINE (FRANCE).

Degani, A., & Heymann, M. (2002). Formal verification of human-automation interaction. *Human Factors, 44*(1), 28. Retrieved from https://search.proquest.com/docview/ 216435950?accountid=15138.

Jordan, Patrick W., B. Thomas, Ian Lyall McClelland, Bernard Weerdmeester. (1996). Usability Evaluation In Industry. *Taylor and Francis*. Retrieved from https://books.google.com/books?id=IfUsRmzAqvEC&printsec=frontcover&source=gbs_ge_summary_r&cad=0#v=onepage&q&f=false

McLeod, Saul. (2017). Qualitative vs. Quantitative Research. *SimplyPsychology.* Retrieved from https://www.simplypsychology.org/qualitative-quantitative.html

Merriam-Webster. (2018). Graphical User Interface Dictionary. *Merriam-Webster Incorporated.* Retrieved from https://www.merriam-webster.com/dictionary/graphical%20user%20interface

Parnell, Gregory S., Driscoll, Patrick J., Henderson, Dale L. (2011). Decision Making In Systems Engineering and Management. Hoboken, NJ: John Wiley & Sons.