Technical Report
1254

Research and Development of Observational Research Tool for Cyber Operations

V.F. Mancuso
S.M. McGuire
P. Picciano
E. Aubin

06 October 2020

Lincoln Laboratory
MASSACHUSETTS INSTITUTE OF TECHNOLOGY
LEXINGTON, MASSACHUSETTS

This material is based upon work supported under Air Force Contract No. FA8702-15-D-0001.

This report is the result of studies performed at Lincoln Laboratory, a federally funded research and development center operated by Massachusetts Institute of Technology. This material is based upon work supported under Air Force Contract No. FA8702-15-D-0001. Any opinions, findings, conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the U.S. Air Force.

Massachusetts Institute of Technology
Lincoln Laboratory

Research and Development of Observational Research Tool for Cyber Operations

V.F. Mancuso

S.M. McGuire

P. Picciano

E. Aubin

Group 57

Technical Report 1254

06 October 2020

Lexington Massachusetts

This page intentionally left blank.

ACKNOWLEDGEMENTS

The work in the following report would not have been possible without the assistance of Diane Staheli, and Gary Ruscinski at MIT Lincoln Laboratory, and direction from COL Stoney Trent, and Lt. COL David Merritt.

This page intentionally left blank.

ABSTRACT

The following report documents a multi-effort research, design, and development of the Behavioral Observations Logging Tool with Surveys (BOLTS). In response to an underlying need from the USCYBERCOM Cyber Immersion Laboratory (CIL), MIT Lincoln Laboratory conducted a user-centered design research project, to understand the needs of the CIL researchers, explore various concepts of operations, and prototype and demonstrate a system at a cyber exercise. This report provides details on the user engagement, concept development, prototyping, and evaluation of the BOLTS system.

This page intentionally left blank.

TABLE OF CONTENTS

This page intentionally left blank.

LIST OF ILLUSTRATIONS

Figure No. **Page**

This page intentionally left blank.

LIST OF TABLES

This page intentionally left blank.

1. INTRODUCTION

1.1 HISTORY AND PROGRAM BACKGROUND

In February 2017, in coordination with the Experimentation Branch at U.S. Cyber Command Capabilities Development Group (USCYBERCOM CDG, now J9), MIT Lincoln Laboratory (MIT LL) stood up a program in support of the Cyber Immersion Laboratory (CIL), entitled "Cyber Human-in-the-Loop Instrumentation and Measures for Assessment". Under the supervision of COL Stoney Trent, MIT LL aimed to help USCYBERCOMAND CDG grow their capabilities in a high fidelity network simulation, human-in-the-loop experimentation and operational research and experimentation. The primary goals of the program were to aid CDG/J9 in fulfilling their key role in experimentation and assessment of cyber teams, technology, and enabling a positive impact on the DoD cyber mission. During the kickoff of the program in March of 2017 two phases were agreed upon, each with three tasks aimed at building technology for measuring individual and team performance:

Phase 1:

1. Observer tablet and analytics-capability for logging observations
2. Application monitor and analytics-capability for capturing features used in web based tools
3. Collaboration monitor and analytics-capability for measuring coordination between team members

Phase 2:

4. Location monitor and analytics-capability to track individual location within a space
5. Gaze tracking-capability for tracking regions of focus on computer screen
6. Physio-behavioral monitor and analytics-capability to measure cognitive workload

The following report details the work conducted for Task 1 between March 2017 and December 2018, in developing the observer tablet and associated analytics. In the next sections, we will provide an overview of the Behavioral Observations Logging Tool with Surveys (BOLTS) highlighting the development of the capability.

1.2 BOLTS OVERVIEW

Behavioral Observations Logging Tool with Surveys (BOLTS) is a flexible, digital observational measurement tool designed as part of a comprehensive approach to assess and improve human-system performance. The work was sponsored by USCYBERCOM in support of the Cyber Immersion Lab (CIL) to:

Conduct evaluation, assessment, and testing of current and emerging technology in support of development, prototyping, acquisition, and operationalization of cyberspace capabilities to meet the operational needs, enhance tactical efficiency, and ensure mission accomplishment of the joint warfighter.

To meet these objectives, we examined a number of research and evaluation efforts conducted in cyber security contexts. This broad assessment exposed several shortcomings and targets for improvement. We then worked to scope the larger applied research vision, decomposing it into elements that fulfill specific needs. Figure 1 shows this larger vision, BOLTS, listed as Observer Tablets in the figure, is part of a larger integrated platform that can produce assessment metrics for candidate solutions for USCYBERCOM. Work continues to mature the suite of capabilities, including physiological measures and application tracking, to round out a comprehensive analysis platform.

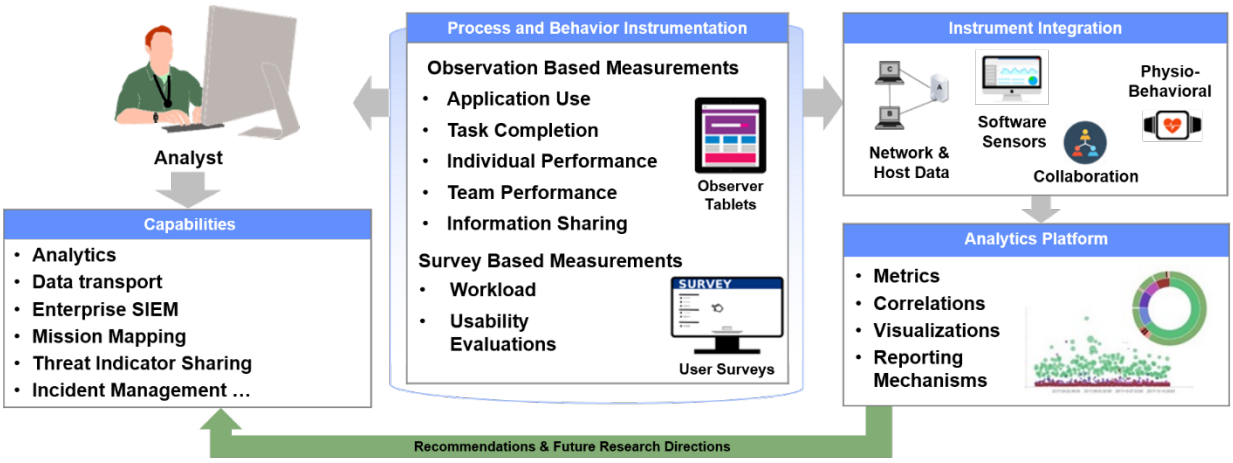

Figure 1: Vision for an integrated platform of diverse data sources to support CIL operational evaluations.

BOLTS provides a means to capture data that are problematic or unfeasible to collect through other channels, such as documenting complex elements of workflows, team coordination, and other activities that reside outside of the digital domain. Another benefit of this collection modality is that it enables embedded observers to produce quantitative and qualitative data through the lens of domain experts.

Researchers and moderators need to not only track events and make notes concerning participant actions and scenario events, but also *how* the scenario is going. Information about the execution of an exercise is critical to manage and improve processes. Prior CIL exercises leaned

on pen and paper as the primary means of data collection and note taking. The functionality of BOLTS aims to reduce the pain and improve the quality of data gathering in support of various data collection needs. Through review of prior CIL exercises and a broad range of user-centered design activities it was determined BOLTS would:

- Enable collection of observer based measurements in a medium that eliminates/reduces the need to capture hand-written data
- Create more accurate timelines of events and processes during an assessment
- Standardize observer based data collection
- Automate processing and analysis
- Facilitate quick after-action reporting
- Deliver survey instruments to observers and users
- Integrate with other sensors and tools

This page intentionally left blank.

2. DEVELOPMENT OF BOLTS

As a basis for our development of the BOLTS, we adopted a user-centered design (ISO 13407, 2010) approach, involving three main processes (Figure 2):

1. Discovery, where we understood the users and current lay of the land,
2. Concept development where we flushed out and iterated on system concepts, and
3. Prototyping and user testing, where we developed a prototype system and evaluated it at multiple fidelities

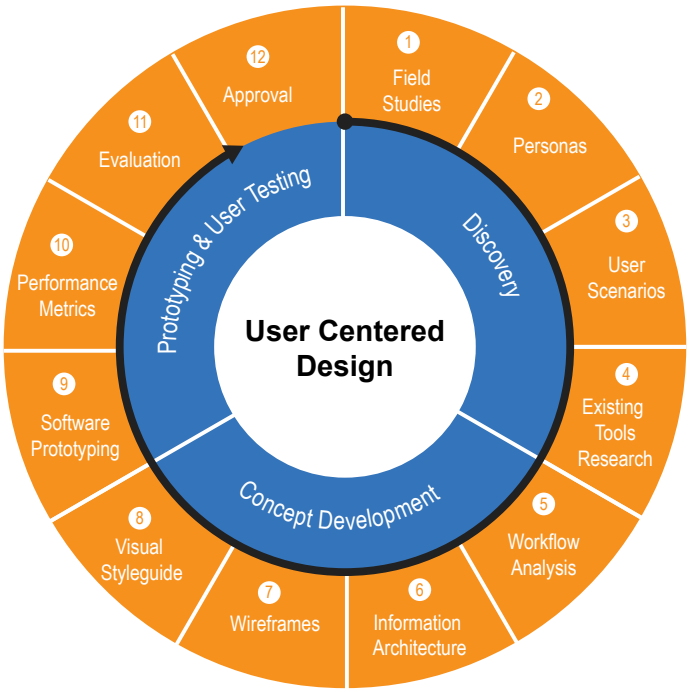

Figure 2: User-centered design process.

A key element of the user-centered design process is the involvement of the target user-base, in this case the CIL research staff, throughout the development life-cycle, from conception through design and development. During the discovery phase, we worked with the users to develop an explicit understanding of their tasks and environment, decision-making processes, job responsibilities and pain points. During concept development we built mockups of potential interface and system design and vetted them with the users. Finally, during prototyping and evaluation, we involved users throughout the entire process. While depicted as a circle, it is important to note that the user-centered design process allows for agility, so as new requirements would come in, we could pivot to prior parts of the process and start from there.

2.1 DISCOVERY

2.1.1 User Requirements

The most salient and targeted needs emerged from a review of an exercise conducted at the CIL. The previous study, was led by internal CIL research staff, and supported by an external contractor, taking place over the course of six months. It was one of the earliest, comprehensive, applied investigations performed in the confines of the CIL. CIL research staff coordinated a massive effort to create infrastructure, and scenarios to execute a high fidelity simulation. They successfully recruited Cyber Protection Team (CPT) members from across the Department of Defense (DoD), and implemented a realistic cyber environment. This confluence of a high fidelity environment, realistic exercises, and representative participants yields immense opportunity to collect highly insightful data, and the output was extremely beneficial in scoping our discovery activities.

To better understand the users and their needs, we conducted structured interviews with study coordinators and research staff. They also provided numerous artifacts from the exercise that enabled a document review of their work. When reviewing artifacts from the data collection, we found an excessive human burden to manually capture performance variables, notes, and information about the state of the scenario. These pain points were echoed in the interviews and discussions with the researchers, who provided further context into the challenges during data collection and analysis.

Throughout the discussions, a few key capabilities consistently emerged as critical for capturing notes and data pertinent for human-in-the-loop observations. The ability to collect consistent data in an efficient manner would substantially aid in:

- Collecting observer based measurements in the laboratory and the field
    - Individual and team behavior
    - Application usage
    - Information sharing
    - Task performance
- Delivering survey instruments to users
- Creating accurate timelines of events during an assessment
- Facilitating quick data analysis for after action reporting

Thus, we set out to improve data collection with a digital tool that could help increase data quality, reduce the potential for errors of commission and omission, and drastically increase the efficiency of analysis by eliminating the need to painstakingly transfer all information from notebooks to computers. However, while it was clear these features could improve the data collection and analysis process, simply seeing the work from a prior exercise was not sufficient. To build out a capable, productive observer tool we explored different types of data collection and analyses scenarios that were likely to be relevant to USCYBERCOM.

Throughout our investigatory process, the use cases continued to point in the direction of an observer. However, we sought to create a flexible platform that could support multiple needs. This included supporting users not only during an exercise, but in other phases of research and activity documentation and reporting.

2.1.2 Existing Tools Research

Based on our findings from the user engagement, the next stage of user centered design was to assess currently available solutions. The goal was not necessarily to see if there was an existing tool that could meet the users' needs immediately, but rather to see if there were solutions that we could either leverage, or learn from, in delivering our final solution. To assess each tool, we came up with a series of domain requirements that we could collect data on and assess the tools utility. Based on the sponsors needs, and the environment that we would have to deploy the final tool, we also came up with a series of limitations that would have to be considered during development:

1. Due to security policy, mobile tablets (i.e., iPad, Android tablets) are not permitted in CIL, so capability must be either delivered as a Windows Application or Web app
2. CIL is a secure facility, and software cannot depend on wireless communication (Wi-Fi, Bluetooth, etc.), so tools should not require internet.
3. All data collected during exercises must be reviewed by security personnel prior to leaving the facility to verify that it does not contain classified data, so only software with local data storage options should be considered
4. Tablets must have stand-alone capability (i.e., does not depend on any other external resources), and should be hosted on internal CIL servers
5. Data must be flexible, and allow for exporting into multiple formats for investigation
6. Tablets will be part of a larger instrumentation ecosystem, so software must be able to be integrated with other COTS and GOTS tools
7. Due to security policy, software must be able to physically disable recording capabilities (i.e., image, audio, and video)

Additionally, we wanted to be sure that whatever solution we select would have the ability to be extended to the specific needs of the CIL research staff and not be cost prohibitive when fully implemented. Table 1 shows an analysis of the more promising solutions we found, analyzed based on a set of domain requirements that could impact their ability to be used in our target environment.

Table 1: List of Existing Tools Reviewed by Domain Requirements: Red Background Indicates Major Issue for Use in the CIL, Yellow Background Indicates Potential Issue

Requirement	Handrail	Fulcrum	ODK	FastField	Device Magic
Platform	Web app	Mobile/Web	Mobile	Mobile/Web	Mobile/Web
Requires Internet	?	No	No	No	No
Local Data Storage	?	?	Yes	Yes	?
Hosting Options	External	External	Local or External	Local or External	External
Export Formats	None	Raw Data	Raw Data	Raw Data	Raw Data
Integration	?	Via API	Via COTS tools	Via COTS tools	Via COTS tools
Video/Audio	?	Yes	Yes	Yes	?
Pricing	Per Month	Per month per user	Free	Per month per user	Per Month per device

In addition to these requirements, we also assessed each tool for user-focused requirements, such as auto-fill, drawing and sketching, customized forms, integrated scoring, etc. While many of the user-focused features met the needs outlined in the previous sections, as we looked deeper, we encountered numerous issues that would undermine the selection of any of the solutions we investigated. Common disqualifiers included:

- Prohibitive cost model
- Built for mobile OS (i.e., Android, iOS)
- Reliance on cloud infrastructure
- Closed systems architecture
- Lack of open API
- Too much functionality/clutter

Most of the tools violated multiple criteria, and they all presented substantial difficulties attempting to make them work in a secure environment. Based on this assessment, we moved forward with developing our own prototype system that could be tailored precisely to the needs of the researchers in the CIL.

2.2 CONCEPT DEVELOPMENT

2.2.1 User Needs Assessment

Using our findings from the user engagement, we set forth in documenting the set of key user needs that would have to be addressed in order to support data collection and analysis (Table 2). This list was compiled based on our own analysis of interviews, and artifacts, as well as matured through direct feedback from the CIL research team. Many of the features were posited in our initial mockups, and proved valid through our alpha version of the software and into the final deliverable.

Table 2: Matrix of User Needs and Features for Observation Support Tool

User Need	Example Forms	Functional Examples
Rapid, structured inputs	Combo box, radio buttons	Rate actions, indicate state or step
Flexible, unstructured inputs	Free text, drawing	Add detail, note anomaly, add reminder
Mark scenario events	Flag, timestamp, categorize, annotate	Mark & label critical points, ID durations of task, go back in scenario or in analyses
Timeline manipulation	Timeline slider	Gives preview of upcoming events, allow user to go back and forth in time during scenario
Simultaneous access	Split screen, tabs, common controls	Timestamp, label, describe event of interest with minimal clicks
Protected access	Require deliberate navigation between previously captured data, current scenario	Don’t want to overwrite data entry from prior run
Prompts to execute investigation events	Pop-ups, alerts/notifications	Remind to conduct Situation Awareness (SA) probe

Table 2, continued

User Need	Example Forms	Functional Examples
Access to investigation resources	Links, tabs	Facilitate capture of post-run surveys
Synch experimental time and UTC	Without connectivity, likely manual with some potential for error	All press start button to begin scenario clocks
Stand-alone functionality and data exchange when connected	Produce text file outputs that can be harvested after run	Produce and organize output files that can be copied

2.2.2 Defining Objectives and Use Cases

Consistent with these needs, we moved toward more tangible guidance to help us decompose this development task. In doing so, we crafted and over time refined, objectives and user stories, all while keeping in mind the firm constraints around which this project would have to navigate in order to be useful.

The overarching objective of this capability is to enable researchers at the CIL to improve and streamline their data collection and evaluation methods. Much of this work was being done with paper and pencil, and while it was sufficient it had some drawbacks, specifically in terms of capturing temporality (i.e., what times did things occur), different types of data, capturing quick observations, maintaining consistent structure across observers, and requiring a highly manual analysis process. Based on this we vetted a series of objectives with CIL leadership and target users:

1. Implement automated process for data collection
2. Accommodate structured and unstructured data
3. Capture temporal data for use in reconstructing timelines of activities
4. Allow for capturing anticipated and unanticipated events
5. Provide structure across researchers for analysis

Based on our objectives, the restrictions of the environment, and careful discussion with the users and sponsors, we developed a set of three primary User Stories that would guide the remainder of the development:

1. As a researcher/observer, I want to be able to log my observations with timestamps so I can create an accurate timeline of my observations during an assessment.
2. As a data manager, I want to be able to export observations from multiple researcher tablets into a single database so I can construct an accurate timeline across all observers for analysis.
3. As a researcher/observer, I want to be able to collect survey data from participants, before, during and after an assessment, so we can collect data on their subjective experiences.

2.2.3 Wireframes

Once we documented our initial use cases and user stories, we made a series of simple wireframes to demonstrate potential functionality that we could implement into our initial prototype. In addition to the user needs, we carefully considered the context of use to inform our design. Specifically, we wanted to make sure that the system was usable in both a touch screen, and laptop computer context. This required consideration of button size, and flexible screen layout that would support both landscape and portrait mode. Starting with the high-level needs, we believed the application would benefit from a modular approach. The components that were determined to be most relevant included:

- **Study Configuration** – Allow researchers to add specific condition information and other meta-data about the study they are going to observe.
- **Observation Module –** Allow researchers to capture predefined observations and see a timeline of what they have observed throughout the session. Additionally, allow for unstructured information based on new observations that may occur.
- **Survey –** Simple surveys that could either be handed to participants or based on observer ratings to capture subjective data throughout a session (i.e., NASA-TLX for capturing workload).

Configuration Observation Survey

Figure 3: Early mockups of the Study Configuration, Observation, and Survey modules.

These mockups were iterated on numerous times with internal research staff at MIT LL who had conducted field observations (not in the CIL, but still relevant to our user stories), as well as CIL research staff who had participated in previous research discussed in the previous section. As we iterated, a more complete picture of what the final capability might look like became clear, and we were able to move into full prototyping mode.

3. BOLTS WALKTHROUGH

The BOLTS software can be run on any Windows tablet or computer. When run on a tablet, screen interaction may be made with a keyboard & mouse/touchpad, stylus, or finger. Careful consideration was made in development to ensure that BOLTS required little time to learn the tool and begin collecting data. An overview of the tool and data files are described in the following sections.

3.1 STUDY DETAILS

The first tab, Study Details (Figure 4), is the only one active when the program is first launched. This is to ensure the necessary parameters are entered prior to the start of data collection including where to store the output files, and which tasks to log.

Figure 4: Primary tabs of the BOLTS tablet.

On the Study Details tab (shown in Figure 5), the observer can enter in information relevant to the data collection. This includes the team they are observing, the task that is being completed, and the experimental session. This information is critical for tracking data, and ensuring that observations to information about the scenario and experimental condition are able to be linked. There are also 3 subtabs on the right of the interface. The first is the Participant's tab; on this tab, the participant's names or id number can be entered along with their role. Role can help in documenting what position an individual held during an exercise; for example, it would be useful to log if an individual was a team lead or an analyst performing tasks.

On the Observations subtab (shown in Figure 6), the details of the tasks that will be logged are entered. These tasks will display as buttons on the Observations tab. The buttons can be grouped together under sub categories. In the example shown in Figure 6, the groupings are labeled as A, B, etc., however names that are more relevant to tasks being logged can be added. Once groupings are defined, the names of the tasks being logged can then be entered. In the example shown, this includes tasks such as *Secure Network Taps* and *Secure Access Points.* For each task, the order the buttons are displayed within the subgroups can also be defined as well as if the task is expected to have a start and end time.

On the Timeline subtab (shown in Figure 7), a task reminder during an event can be entered. For example, an observer might want a reminder when they need to have participants complete a survey. For each event, the time from the start of the exercise that it will occur needs

to also be entered. For a reminder on what the event is referring to, a brief description can also be added.

While all of the information just described can be entered manually within BOLTS, a JSON file can also be loaded using the *Load Settings File* button. Loading a file is a quick way to add prior saved files and to make sure the same settings are loaded across multiple observers. Once all information has been entered or uploaded using a file, the *Initialize Scenario* button can be pressed to begin data collection.

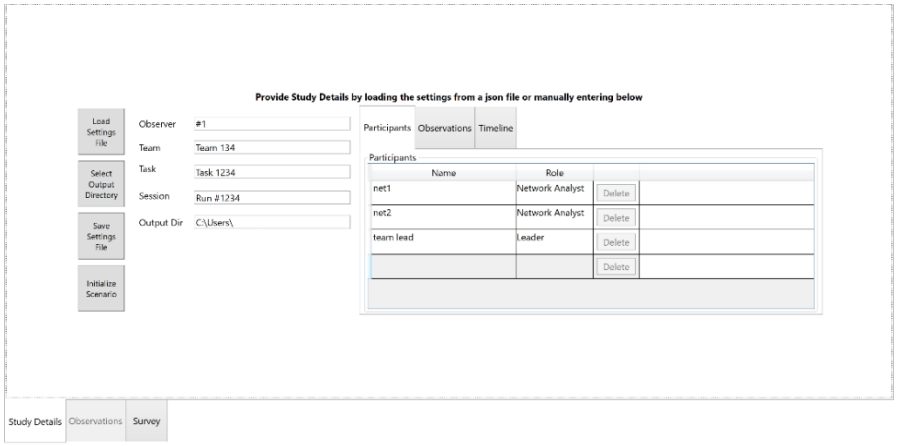

Figure 5: Study Details tab of the BOLTS tablet. On the right the Participants tab is selected.

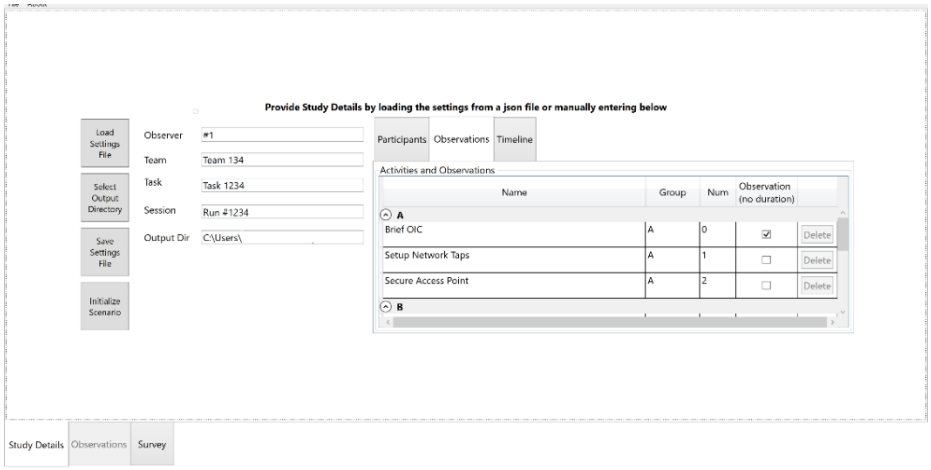

Figure 6: Study Details tab of the BOLTS tablet. On the right the Observations tab is selected.

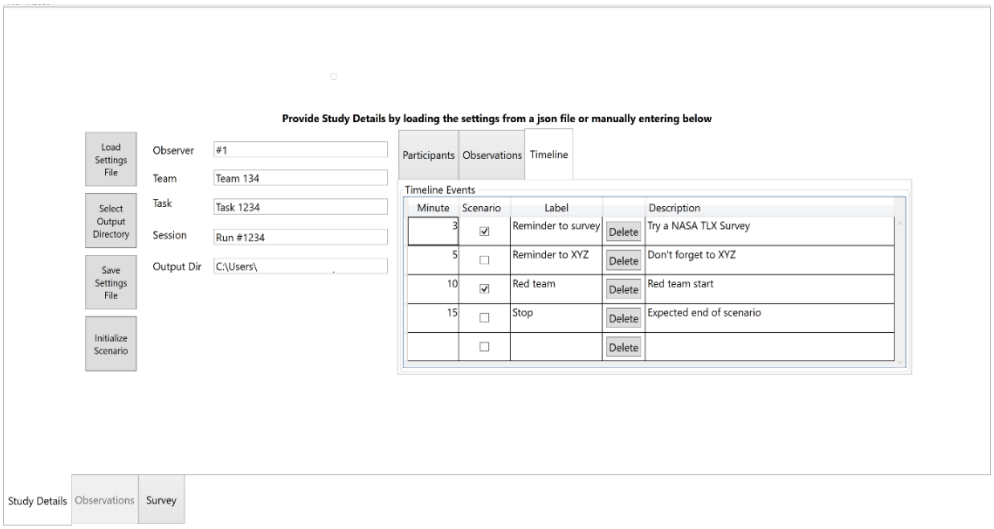

Figure 7: Study Details tab of the BOLTS tablet. On the right the Timeline tab is selected.

3.2 OBSERVATIONS

The second tab of BOLTS is Observations, which is where the observer will be able to log their observations during an experiment, shown in Figure 8. At the beginning of the data collection, the *Start Session* button in the upper left hand corner should be pressed. This allows observers to set a specific and coordinated start time for the scenario. Note that the application will record timestamps even if this button is not pressed first, however this starts the elapsed time at the top of the display. At the top is also the timeline of key events that the observer needs to be aware of. The displayed *Time Until* will change with elapsed time to provide continuous situational awareness of upcoming events.

To the right of the interface is the activity and observation buttons organized within subgroups. Observation buttons are shown in grey and are defined as events which occur at a single point in time and do not require tracking their duration. Activity buttons are shown in yellow and are tasks for which there is an interest in collecting a start time and an end time. For example, an event could be something like requesting information from the White Cell, where you only need to know when it occurred, while an Activity could be something like a team meeting, where you want to know when it starts, and ends and what events occur within the timeframe. When an activity button is pressed, a button also appears in the *Activities to Be Completed* section of the interface. Pressing this button allows the observer to indicate that the activity has completed.

When buttons are pressed, the time of the event is logged and the event appears in the timeline on the interface, therefore there is a running sequence of events that is captured. For each event, there is an *Edit* button. When clicked an observer can cross-out an event if it was entered by mistake, add a note to provide more context to an event, or add the name or id of the

participant that completed the task. We made the strategic decision to not allow researchers to delete events, only allowing them to "cross-out," while allowing the Principal Investigator to make a final determination on which data to remove from analyses. This also helps prevent accidently deleting data. The files of the logged observations are saved in a comma separated file, making it quick to pivot after an event from data capture to data analysis.

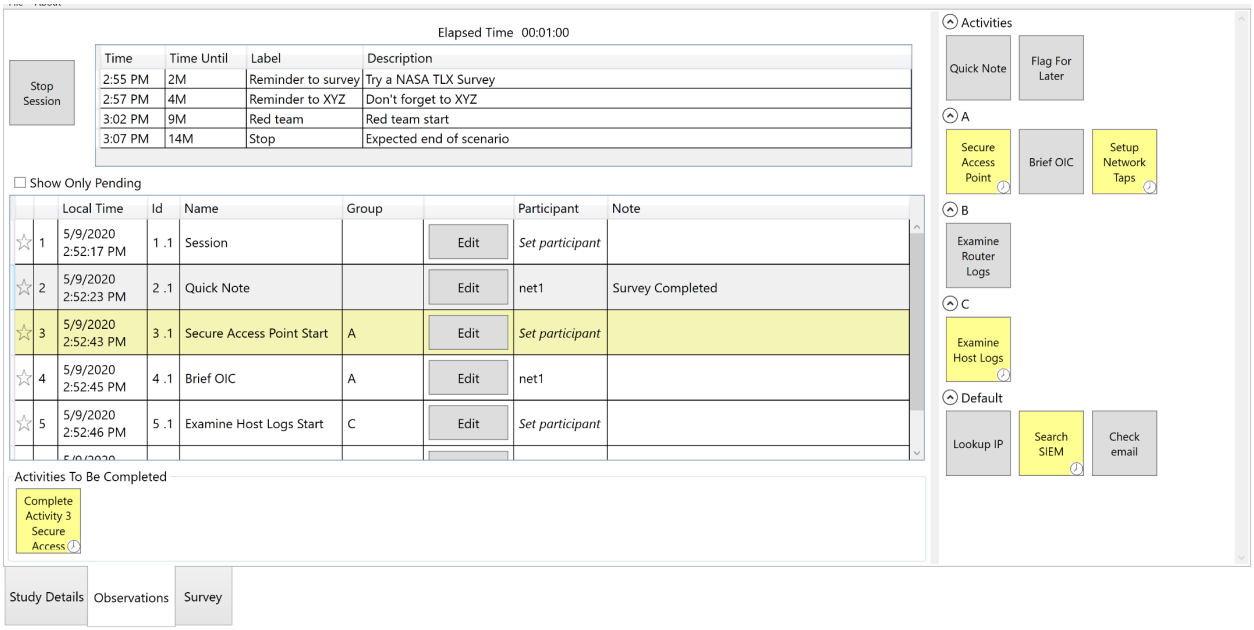

Figure 8: Observation tab of the BOLTS tablet.

BOLTS also uses a dynamic ID system that allows capture of the temporal order of occurrences, while also being able to link together captured events. For example, if you had an activity that started and it was given the ID 4.1, the stop action would be given the ID 4.2. Similarly, any note you add to an activity or observation, or if you cross an Activity or Observation out, would use the major number and add on after the decimal point. This allows us to maintain an accurate timeline, while also capturing subsequent data.

3.3 SURVEY

The third tab of BOLTS is the *Survey* tab. There is a built in version of the NASA Task Load Index (NASA TLX; Hart, 2006), which is a survey of workload. It is shown in Figure 9, and contains sliding scales for each question. At the bottom of the survey, the observer can select the participant that is completing the survey.

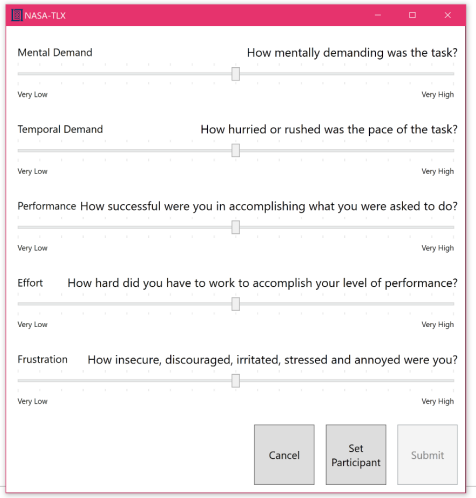

Figure 9: NASA TLX Survey.

In addition to NASA-TLX we have also implemented the Observation Assessment of Teamwork (OAT) survey, which is specifically developed for capturing subjective information about teamwork in cyber security competitions (Buchler et al., 2018). In addition to prepopulated surveys, we have also incorporated functionality to add custom surveys. An example of question types that can be built into a survey are shown in Figure 10. Sliding scales can be added as well as short answer type questions. The surveys can be created in JSON format and uploaded through the Study Details page. The purpose of this custom survey tool was not to build a replacement for commonly used survey capabilities like Qualtrics, SurveyMonkey, and Lime Survey. Rather it was to provide an ability to quickly capture answers to a few important questions during an exercise.

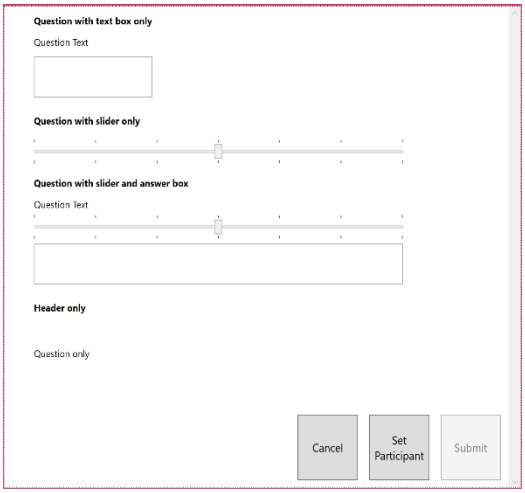

Figure 10: Custom surveys.

This page intentionally left blank.

4. EVALUATIONS

4.1 HEURISTIC EVALUATION

Once we had an initial functioning prototype of BOLTS, a heuristic evaluation was conducted by internal experts in human factors at MIT LL. Nielsen (1992) defined ten heuristics that a system should meet to minimize usability problems. One heuristic is the *visibility of system status*, in which a system should be built in a way that it provides appropriate feedback to the user. One way the BOLTS meets this heuristic is that it provides an event timeline. When a user presses a button activities and observations appear in the timeline allowing them to refer back to the data that they captured. Another heuristic is *flexibility and efficiency of use.* It is important to design tools that can be used at all experience levels. BOLTS provides multiple ways to add information on participants and events to log including manually using the interface, or a more experienced user could create and upload JSON files. An additional heuristic is *help user recognize, recover from errors.* A user is never going to use a system without making any mistakes. Within BOLTS, when errors are made when logging observations, they can be crossed out. While all ten heuristics are not highlighted in this report, BOLTS was designed and built to meet these fundamental usability heuristics to ensure that it would be easy for any individual to use it for collecting observations.

4.2 USABILITY EVALUATION

Following the heuristic evaluation, we expanded on our evaluation by having individuals unfamiliar with BOLTS learn and use it and provide feedback on its usability in a laboratory task. For the evaluation, we constructed a simple two person task that consisted of having users work together to complete a small puzzle. This was a good task for novices to complete and record their observations as it required less than 30 minutes, there were clear performance indicators, and it required teamwork between the two completing the puzzle as they had to communicate and share pieces.

For the evaluation, we recruited two participants who had experience in conducting user observations, however were not familiar with BOLTS or the CIL use cases we used to develop it. During the evaluation, participants were asked to log activities and observations using BOLTS. They were asked to log when a corner piece was set, when person A was requesting information from person B (and vice versa) or when giving direction, and when sections of the puzzle were completed and combined. To become acquainted with BOLTS they were provided a brief orientation that took approximately five minutes. Both observers were able to quickly become familiar with and utilize the BOLTS interface. In addition, no evaluation session had to be halted for technical difficulties or failures on the part of the observers to understand how to use the tool. The duration of activities logged by both observers is listed in Table 3. We found that the two observers were able to consistently record data. In the table the numbers within the columns for Observer 1 and Observer 2 represent the duration (in minutes) each observer recorded for task

completion time. For most activities, which were completing quadrants of a puzzle, the observers logged duration times that were within a minute of each other.

Table 3: Duration of Logged Activities (minutes)

Action	Observer 1	Observer 2
Puzzle 1 Top Left Section Completed	3.4	3.6
Puzzle 1 Top Right Section Completed	6.4	4.0
Puzzle 1 Bottom Left Section Completed	2.6	2.2
Puzzle 1 Bottom Right Section Completed	2.3	2.3
Puzzle 2 Top Left Section Completed	4.9	2.5
Puzzle 2 Top Right Section Completed	2.1	1.9
Puzzle 2 Bottom Left Section Completed	2.8	2.7
Puzzle 2 Bottom Right Section Completed	2.0	Incorrectly Logged

From this evaluation we also wanted to determine if the data could be utilized for evaluating performance. Figure 11 shows a plot of the tasks against the duration of the scenario (minutes). The circle size indicates a greater number of occurrences of logged events in a given minute. The lines of tasks represent a continuous duration. Using these analyses, we can compare teams by completion success, task duration, and examine workflow (timing and sequence of tasks), which can be useful in characterizing performance.

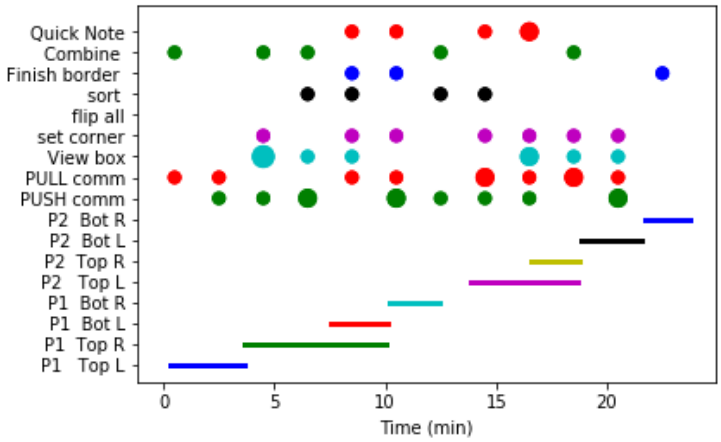

Figure 11: Observations captured over the course of the pilot scenario (tasks vs. time).

The puzzle pilot afforded not only the opportunity to obtain feedback on the BOLTS interface from unfamiliar participants, but a chance to use it in an active setting. Deploying the tool to perform task-driven observations drove a number of recommendations to facilitate improved usage in a live setting. In addition, we were able to establish analysis methods for analyzing data post collection.

4.3 OPERATIONAL EVALUATION

To evaluate the use of BOLTS in an operational setting, we attended a five-day cyber exercise located at the CIL. The exercise consisted of two CPT teams and the purpose of the exercise was to identify malicious activity on a network. During the study, three members of the MIT LL research team collected data (one per team, with another who floated depending on which CPT was surging). Additionally, during one of the days, we recruited a member of the CIL research staff to use BOLTS to collect observations of the CPTs. Like participants in the initial evaluation, the CIL researcher had no familiarity with BOLTS, and received a short five minute overview of how to use it, and was able to quickly get up to speed.

On the first day of the exercise, we familiarized ourselves with the teams, range, and other attendees. On day two we setup and configured the tablets, installing the software, and configuring information about the scenario, and prepopulating the Observations and Activities we wanted to capture. We then logged observations for two days and spent the final day reviewing the data with security personnel and transferring the data to media to remove from the CIL.

During the exercise our observations were focused on information sharing. The full list of activities and observations that were logged are listed in Table 4 and Table 5. Activities are events in which we wanted to capture the duration of the event. For example, we wanted to document the time it took from when a request was made to the white cell until the team received

an answer. We also wanted to document when expertise was shared between members of the team. Observations are events in which we wanted to document that an event occurred but not the duration. This included when a discovery was made, when a specific tool was used, or when there was a strategy shift of the team in how they tackled the problem. In addition to logging activities and observations we also recorded typed notes to document tool and resource names, specifics about what was discovered among other types of information.

Table 4: Activities Logged During the Cyber Exercise

Activities	Description
Request: White Cell	Request made by team for information
Inform Team	Information was shared with all CPTs on a team
Expertise Sharing	An individual sharing tool use or hunt knowledge

Table 5: Observations Logged During the Cyber Exercise

Observations	Description
Coordination	Two or more CPTs working together
IoC Discovery	A potential IoC was found
Utilize Resource	A reference such as a book, poster, website, etc. was used
Strategy Shift	A shift in the current hunt approach
CMD Info Request	Team lead requesting information
Tools	Log of tools being used by CPTs with notes on their activities within the tools

In addition to data on the exercise, we also created issue logging tickets, which could be filled out if the research team experienced bugs, issues or ideas that could be implemented in the future.

Results from the exercise are shown in Figure 12. The captured observations were saved in an easy to use comma separated value (csv) file, which allowed for a quick turnaround between data captured and analysis of the activities during the exercise. Based on the number of observations captured, we examined the operations (ops) tempo of the teams. We found that both teams had an increase in activity at the start of the exercise followed by lower levels of

activity throughout the day. We also noticed that overall there was more coordination activity between team 1 than team 2 especially on the second day.

In addition to capturing overall activity across the two teams, we also were able to capture coordination by analyst type. We logged events by host and network analysts and those coordination activities that were initiated by the team lead. As shown in Figure 12, the two teams had different coordination strategies. In team 1, coordination was mostly within groups of analysts, while in team 2, the coordination was led by the team lead. The analysis of the collected data demonstrates how BOLTS can be used to collect meaningful data on team performance.

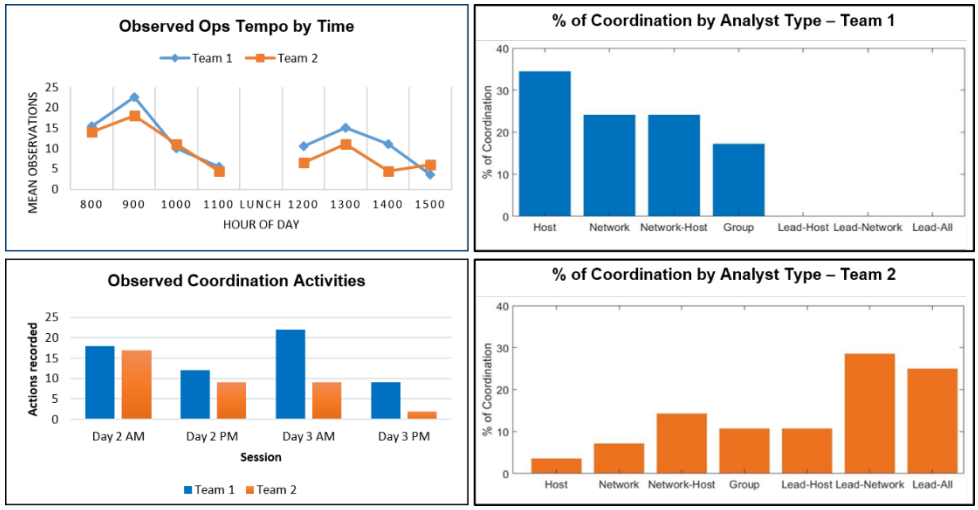

Figure 12. Operations tempo during the exercise and coordination between sub-teams captured using BOLTS.

The exercise provided an opportunity to test our ability to record data across multiple observers and multiple days. We found that data was able to be logged in a consistent manner across all MIT LL observers and the CIL research staff member, resulting in meaningful results on coordination. The exercise was also used as an opportunity to improve on the BOLTS. We obtained feedback from other researchers at the exercise, as well as observers on our team identified a list of features that could be further improved based on using BOLTS across multiple days. Many of the items identified were prioritized and integrated shortly after the exercise to further improve logging of observations.

This page intentionally left blank.

5. FUTURE DEVELOPMENT

BOLTS has gone through several rounds of testing and evaluation in order to build an effective tool for logging observations. There are further areas of development though that could be conducted to expand the capability. In the Evaluations section, the analysis shown was completed post data collection. The observation files were downloaded and analyzed using Python scripts after the events. During an exercise though it would be useful to be able to quickly provide statistics within the BOLTS software or in a co-located auxiliary tool. This would be useful to a researcher who may want to check the data being collected. Also, this would be useful to the individuals or teams themselves, as a review of their performance could be made as part of an after action review. Therefore, an analytics platform is envisioned as being part of a future state of BOLTS. Mockups of the platform are shown in Figure 13. The concept is to provide a flexible tool that allows many views of the data to be generated quickly and easily by the observers. An observer could select the team or participant, select the statistics, and then the results can be viewed and compared.

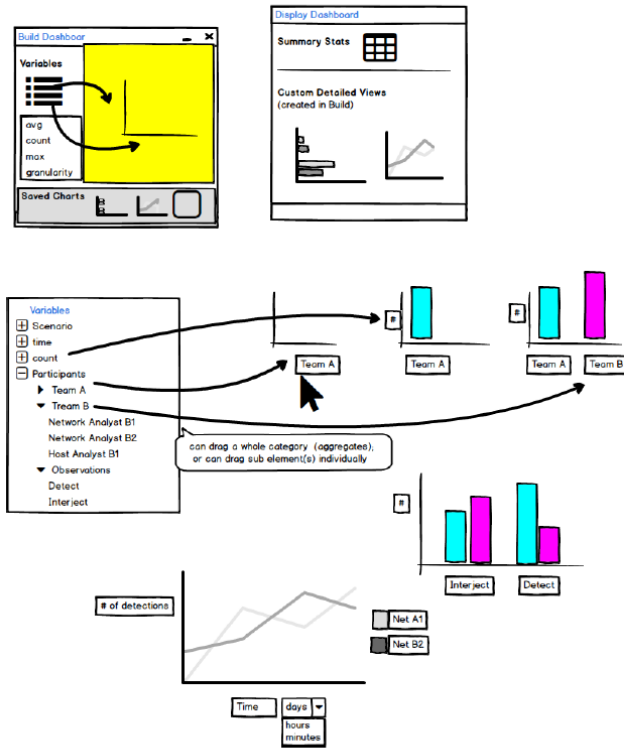

Figure 13: Mockups of data analysis platform for BOLTS.

The mockups shown on the analysis platform is for a single tablet. There is also value in aggregating the data across teams and observers in order to summarize the data. This could be particularly useful for hotwashes conducted at the end of the exercise. Therefore, it could be useful, when the tablets can be connected to a network, to have a capability in which the data could be collected from multiple tablets and aggregated on a common server automatically. Analysis could then be performed, resulting in a comparison of observation data across individuals or teams and summarized in a web based user interface.

BOLTS is one instrument for obtaining measurements on individual or team performance. However, there are many other instruments that could be used as shown in the larger vision in Figure 1. It would therefore be useful to integrate BOLTS data into a larger platform for capturing measurements of individual performance, team interaction, and physiological sensors that can capture data that can be used to assess workload. MIT LL has developed a platform that integrates a few instruments including application monitoring, eye tracking, and heart rate measurements. A future step would be to incorporate BOLTS data into this platform's analytics capability.

6. TRANSITIONS AND DEPLOYMENTS

Since its initial demonstration in the CIL for the Hunt Competition, BOLTS has been successfully transitioned to numerous sponsors, and has been used across multiple use cases. In addition to the software, in many of the situations we successfully transitioned the entire codebase which was adapted to add features that were specific to the organizational mission and/or project. Below is a summative list of successful deployments of the BOLTS system:

1. Evaluations of Cyber AI technologies in a team task to help inform accreditations and acquisitions for use on DoD systems
2. Observations collegiate cyber security competitions such as the National Collegiate Cyber Defense Competition and U.S. Cyber Challenge: Cyber Quest to inform the development of measures of performance in cyber environments.
3. Comparison of user performance for completing tasks using a work aid with a baseline/manual condition to inform the development of a COP for CCMD CPTs.
4. Assessment of functionality and user experience of a prototype system during a usability evaluation
5. Augmenting quantitative data collected by a system, and providing a more comprehensive look at the tactics, preferences, characteristics, and tendencies of the human operator while interacting with an AI agent.

In addition to our successful transitions, MIT LL has continued to explore opportunities to leverage BOLTS for other sponsors and domains. Currently we have two potential opportunities to use BOLTS for future research focused on better understanding processes of a Red team and CPT during training exercises.

This page intentionally left blank.

7. SUMMARY AND CONCLUSION

BOLTS was developed to provide a flexible tool for capturing observations. By adopting a user-centered design approach we were able to develop a tool that meets user needs. Through multiple evaluations we were able to iterate and refine on the design. The result is a tool that requires minimal time to learn, can be utilized consistently across observers, and can be setup to log observations for a range of events. This flexibility and ease of use has resulted in BOLTS being used for a wide range of activities and be transitioned to a range of sponsors for use in documenting workflows, team coordination, and evaluation of tools.

Measurement of performance in cyber operations is a critical need for evaluating the efficacy of operations, tools and training. Currently, much of the performance-based assessments in cyber rely on subjective evaluations by subject matter experts, and outcome-based measures such as “did they find the attacker”. While these provide insight into higher level processes and overall performance, we need tools that help researchers collect and analyze data that provide insights into more fine-grained processes including tools use, information sharing, communication and coordination. BOLTS was the first tool of the larger envisioned instrumentation suite being developed at MIT LL to provide such an essential capability.

This page intentionally left blank.

REFERENCES

Buchler, N., Rajivan, P., Marusich, L. R., Lightner, L., & Gonzalez, C. (2018). Sociometrics and observational assessment of teaming and leadership in a cyber security defense competition. computers & security, 73, 114-136.

Device Magic [Computer software]. (2011). Retrieved from http://www.devicemagic.com

Fastfield [Computer software]. (2016). Retrieved from http://www.fastfieldforms.com

Fulcrum [Computer software]. (2014). Retrieved from http://www.fulcrumapp.com

Handrail [Computer software]. (2015). Retrieved from http://www.handrailux.com

Hart, S. G. (2006, October). NASA-task load index (NASA-TLX); 20 years later. In proceedings of the human factors and ergonomics society annual meeting (Vol. 50, No. 9, pp. 904-908). Sage CA: Los Angeles, CA: Sage Publications.

ISO 13407, Human-centered design processes for interactive systems. 2010, ISO: Geneva.

Nielsen, J. (1992). Finding usability problems through heuristic evaluation. Proc. ACM CHI'92 (Monterey, CA, 3-7 May), pp. 373-380.

Open Data Kit (ODK) [Computer Software]. (2008). Retrieved from http://www.opendatakit.org

This page intentionally left blank.

REPORT DOCUMENTATION PAGE	Form Approved OMB No. 0704-0188

PLEASE DO NOT RETURN YOUR FORM TO THE ABOVE ADDRESS.

1. REPORT DATE *(DD-MM-YYYY)*	2. REPORT TYPE	3. DATES COVERED *(From - To)*
06-10-2020	Technical Report	

4. TITLE AND SUBTITLE

Research and Development of Observational Research Tool for Cyber Operations

5a. CONTRACT NUMBER

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)

Vincent F. Mancuso, Sarah McGuire, Paul Picciano, and Ethan Aubin

5d. PROJECT NUMBER

2767

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

Lincoln Laboratory, MIT
244 Wood Street
Lexington, MA 02421-6426

8. PERFORMING ORGANIZATION REPORT NUMBER

TR-1254

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)

US Cyber Command
Suite 6171
9800 Savage Road
Fort George G. Meade, MD 20755

10. SPONSOR/MONITOR'S ACRONYM(S)

USCYBERCOM J9

11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT

Distribution Statement A

13. SUPPLEMENTARY NOTES

None

14. ABSTRACT

The following report documents a multi-effort research, design, and development of the Behavioral Observations Logging Tool with Surveys (BOLTS). In response to an underlying need from the USCYBERCOM Cyber Immersion Laboratory (CIL), MIT Lincoln Laboratory conducted a user-centered design research project, to understand the needs of the CIL researchers, explore various concepts of operations, and prototype and demonstrate a system at a cyber exercise. This report provides details on the user engagement, concept development, prototyping, and evaluation of the BOLTS system.

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			19b. TELEPHONE NUMBER *(Include area code)*
U	U	U	U	44	

This page intentionally left blank.