The Tularosa Study: An Experimental Design and Implementation to Quantify the Effectiveness of Cyber Deception

Kimberly J. Ferguson-Walter,
Temmie B. Shade, Andrew V. Rogers
U.S. Department of Defense
kimberly.ferguson-wa@navy.mil

Elizabeth M. Niedbala
Texas Tech University
elizabeth.niedbala@ttu.edu

Michael C. Trumbo, Kevin Nauer,
Kristin M. Divis, Aaron P. Jones,
Angela Combs, Robert G. Abbott
Sandia National Laboratories*
rgabbot@sandia.gov

Abstract

The Tularosa study was designed to understand how defensive deception–including both cyber and psychological–affects cyber attackers. Over 130 red teamers participated in a network penetration task over two days in which we controlled both the presence of and explicit mention of deceptive defensive techniques. To our knowledge, this represents the largest study of its kind ever conducted on a professional red team population. The design was conducted with a battery of questionnaires (e.g., experience, personality, etc.) and cognitive tasks (e.g., fluid intelligence, working memory, etc.), allowing for the characterization of a "typical" red teamer, as well as physiological measures (e.g., galvanic skin response, heart rate, etc.) to be correlated with the cyber events. This paper focuses on the design, implementation, data, population characteristics, and begins to examine preliminary results.

1. Introduction

Deception is the provision of misinformation that is realistic enough to confuse situational awareness and to influence and misdirect perceptions and decision processes. Deception-based cybersecurity techniques offer potential advantages over traditional security controls, and computer networks are natural settings for inducing confusion due to their inherent complexity and the uncertainty in accessing and understanding them remotely. Though offense has traditionally had the upper hand in the cyber realm, cyber deception is an emerging area of research aimed at rebalancing this asymmetric nature of cyber defense [1, 2].

Cyber deception is potentially a powerful defensive tactic because instead of just affecting an attacker's behavior (by temporarily blocking access), it also affects their decision-making processes, causing them to waste time and effort based on incorrect information. However, there is little experimental evidence of how effective this kind of defense can be or how it may compare to other defenses. In order to measure the (positive) impact on cyber defenders and the (negative) impact on cyber attackers, we must use interdisciplinary research teams that combine cyber security with behavioral science experts. This paper discusses the rigorous experimental design and protocol we used and the data we collected in the Tularosa study. We discuss preliminary results addressing one of our research hypotheses and our plan for future data analysis to address the remaining.

We designed a network penetration testing exercise on a simulated computer network with different conditions examining how the effect changes if a cyber deception technique is used or not and whether the participant is aware of the deception or not. We also sought to understand the attackers' cognitive, emotional, and physiological responses to the deception, which is the human subjects research (HSR) portion of the study. This information was collected to better understand when deception is effective and better correlate the impact on the human (i.e., stress, confusion, frustration) with cyber task performance.

The research study design used a range of personality indices, physiological measures, and cognitive tests to understand attackers' mental models, decisions, and behaviors. Each index, measure, and test was hypothesized to correlate with performance on the network penetration task in the presence of deceptive defenses. However, while prior work has conducted qualitative interviews with individuals who engage in red team activities [3] and others have suggested cognitive processes that may predict success in network penetration attempts [4], these previous research efforts stopped short of collecting data in order to obtain an objective characterization of cognitive abilities and personality attributes. This can limit our ability to predict the effectiveness of proposed defenses and to identify and utilize opportunities to thwart attacks. In many cases, our cognitive tests do not correspond to information that is directly available about attackers on an individual basis in a real-world setting. However, understanding the correlations between these factors

* Sandia National Laboratories is a multi-mission laboratory managed and operated by National Technology and Engineering Solutions of Sandia, LLC., a wholly owned subsidiary of Honeywell International, Inc., for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-NA0003525 SAND2018-854142.

URI: https://hdl.handle.net/10125/60164
ISBN: 978-0-9981331-2-6

and task performance may allow us to categorize attackers and mount a tailored response.

2. Related Work

One common deception technique is the use of honeypots to lure, contain, and observe attackers and their activities [5]. This technique has been explored and expanded upon in many ways. There are technologies that present an adversary with a false network topology [6]. Others [7] have created a framework for deception to assist in the analysis of deceptions, whether it involves people or computers, networks of people or computers, or people paired with computers. Honeypot effectiveness has been tested using cyber security games, revealing how different setups may cause attackers to change their operations to avoid negative outcomes [8]. Based on the success of honeypots, fake honeypots were developed to make real systems look fake to deter attacks [9].

Decoy systems similarly use deception techniques but differ from honeypot technology [10]. Decoy systems are typically embedded within the true network, can be configured to make homogenous networks look more heterogeneous, and are low-fidelity, requiring less maintenance than full honeypots [11]. The main purpose of decoys is to obfuscate the network and confuse the attacker about the true network topology.

To date, little work has been done to rigorously prove the efficacy of decoy systems. Few experiments have been executed, and none of these have had rigorous experimental control and/or a large enough sample size of participants that generalize to the desired population. Participants in studies with larger participant pools typically either use unknown parties from the internet [12, 13, 14] or students from universities pursuing cyber-related degrees [9, 15]. Neither of these participant pools generalizes well for predictive results of sophisticated cyber attackers. Placing deception on internet-facing network nodes does attain adversarial activity, but it is uncontrolled, uncertain, and does not allow for insight into the participants by way of reports or interviews. Students lack the experience and mindsets that would parallel the sophisticated adversaries these defenses are employed to deceive. In our study, we look to address these issues by utilizing the closest analogous group to malicious cyber adversaries available for scientific testing–red teams–and bringing in a large enough number of them to guarantee the proper statistical power and reliability to detect effects.

3. Design

Subjects participated in a network penetration task. Modeled after the pilot studies in [11], cyber deception (decoy system present vs. decoy system absent) and explicit mention of deception (informed vs. not informed) were modified between participants on the first day of testing. They were also controlled within participants between the first and second days of testing. Figure 1 contains further details on these conditions. Participants were pseudo-randomly assigned to one of the four conditions. Participants were run in cohorts of varying numbers with all members of a given cohort assigned to the same condition when possible, but some cohorts had mixed conditions in order to achieve a balanced number of participants between conditions. In addition, participants completed a series of questionnaires and cognitive tasks (see Section 3.2).

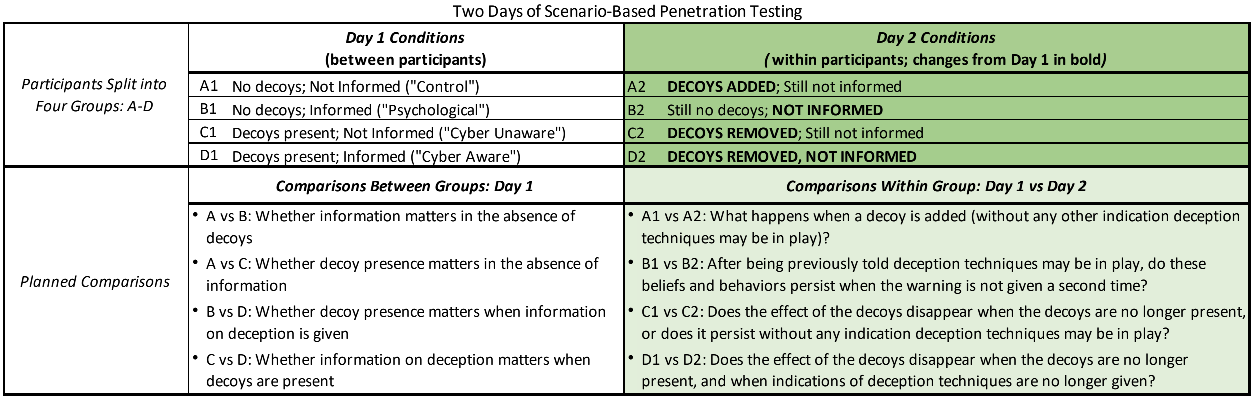
3.1. Cyber Range

The Cyber Range was designed to emulate a semi-realistic enterprise network. Each participant's environment was also instrumented so that data could be collected during the study yet not be intrusive for the participants. The simulated enterprise environment consisted of a number of servers and workstations running both Microsoft Windows and Linux. A wide spectrum of operating system versions were installed.

Active Directory services were installed on a Windows Server 2016 Enterprise system to emulate a typical corporate controlled authentication system. A DNS was also installed to provide name services for all of the clients and servers in the network. There were also a number of other common services provided on the network such as web servers, database servers, file servers, etc. A total of 50 systems were installed in the environment with an even split of 25 each for each operating system–Windows and Linux. Twelve of the 50 were servers (6 Windows servers and 6 Linux servers). In the cyber deception conditions (C1, D1, and A2), there were 50 decoys in addition to these systems.

A Network Time Protocol (NTP) server was configured and installed within each participant's environment to provide time synchronization of all of the machines, allowing for reliable timestamping for data collection. The NTP server was designated as out of scope for the participants since this was part of the experimental support infrastructure and would jeopardize the data collection effort if attacked.

To provide a more realistic environment, 130 domain user accounts were created to provide a lived-in network appearance for the participants. Of the 130 domain user accounts, there were 15 domain administrators to simulate the IT staff for the fictional organization. Separate OUs (Organizational Units) were created to simulate actual business organizations (i.e. IT, Sales, HR, Staff) since it is common to create OUs to mirror primary business functions. A number of files were also created and placed in the user accounts and log files were populated by having staff members perform activity on the network during the creation phase. For example, several domain administrator accounts were selected and used during this phase to generate



Two Days of Scenario-Based Penetration Testing

	Day 1 Conditions **(between participants)**	*Day 2 Conditions* ***(*within participants; changes from Day 1 in bold*)***
Participants Split into Four Groups: A-D	A1 No decoys; Not Informed ("Control")	A2 **DECOYS ADDED**; Still not informed
	B1 No decoys; Informed ("Psychological")	B2 Still no decoys; **NOT INFORMED**
	C1 Decoys present; Not Informed ("Cyber Unaware")	C2 **DECOYS REMOVED**; Still not informed
	D1 Decoys present; Informed ("Cyber Aware")	D2 **DECOYS REMOVED, NOT INFORMED**
Planned Comparisons	***Comparisons Between Groups: Day 1***	***Comparisons Within Group: Day 1 vs Day 2***
	• A vs B: Whether information matters in the absence of decoys	• A1 vs A2: What happens when a decoy is added (without any other indication deception techniques may be in play)?
	• A vs C: Whether decoy presence matters in the absence of information	• B1 vs B2: After being previously told deception techniques may be in play, do these beliefs and behaviors persist when the warning is not given a second time?
	• B vs D: Whether decoy presence matters when information on deception is given	• C1 vs C2: Does the effect of the decoys disappear when the decoys are no longer present, or does it persist without any indication deception techniques may be in play?
	• C vs D: Whether information on deception matters when decoys are present	• D1 vs D2: Does the effect of the decoys disappear when the decoys are no longer present, and when indications of deception techniques are no longer given?

Figure 1. Experimental conditions. Each day, a decoy system was either present or absent on the network and participants were either informed or not informed that cyber deception tools might be present on the network.

log activity that would appear to be authentic. Each domain user account had a password that met the minimum-security requirements. Each Windows system was then joined to the domain to allow authentication services to work properly within the network. The IP addresses were randomized within a typical Class C subnet and MAC addresses were generated to represent typical vendor (i.e. Dell, HP, Intel, etc).

The decoys used as the cyber deception component in this experiment were based on lightweight virtualization and were configured to replicate operating system and services of typical assets residing in an enterprise network. The decoys were configured to mimic both Linux and Windows services similar to those in the Cyber Range. These decoys respond to typical network port scans and provide almost identical feedback to those of real desktops. Since actual services were not running on these decoys, any attempt by a subject to logon failed and was logged as an indicator of unauthorized activity. Examples of some of the services are Apache web server, DNS, SSH, and FTP. Separate environments were designed to facilitate experimental conditions with and without decoys. With exception of the presence or absence of decoys, the environments were designed to be as similar as possible to allow for easy comparative performance analysis.

Each subject was provided with a laptop to use during the experiment that was connected to the cyber range via a dedicated network. These laptops were configured with Kali Linux which provides a robust environment for penetration testers with over 600 security-related tools. Some of the most commonly used tools in this distribution are Nmap (port scanner), Metasploit Framework (penetration testing), and ZAP web application security scanner. In addition, the laptops were configured with their own offline Kali Linux repository with 65 Gigabytes (GB) of binary packages that include additional tools and software that could be easily installed by the subjects. The use of the offline repository enabled us to disconnect the laptops from the internet (ensuring no PII was accidentally collected) while still enabling the subject to install additional software if needed throughout the study.

Cyber Data. We collected several data sources from the participants' attack clients during the study. Netflow and tcpdump recorded full packet capture from their machines for post-experiment review of their network activity. A keylogger and video screen capture were used for the duration of the experiment to record their host-based operations. Participants were encouraged to keep a running log of findings via a Mattermost chat client during the experiment, giving real-time insight into what parts of their activities they thought were notable as they experienced them. Additionally, we retrieved data from the participants' laptops after the experiment was over. Several logs from the Kali Linux operating system were collected, including logs of the processes run, the system notifications, daemon logs, authentication records, and default package logs. The shells used by the participants had their history aggregated to reveal commands entered. All notes stored by the participants on the attack client were collected as well. If deception was present in their environment, we also collected the logs server-side from the decoy system that tracked instances of the decoys being triggered. These logs tracked four primary interactions with the decoys: single packets to a single host (touch), multiple packets to a single host (probe), single packets to multiple hosts in succession (scan), and interactive login attempts (intrusion).

3.2. Individual Measures

In addition to the network penetration task, participants completed a series of questionnaires and cognitive tasks. This section highlights the tasks selected and justification for their inclusion. See supplemental materials in the online appendices[1] for more details including the cyber task instructions (Appendix B), full schedule (Appendix C), and all questionnaires.

Task-Specific Questions. Three sets of questions

[1] https://cfwebprod.sandia.gov/cfdocs/CompResearch/docs/TularosaAppendix.pdf

were designed to measure participants' experiences during the experiment. The questions provided a data stream on task performance in addition to data collected directly during the network penetration task. These questions included a daily briefing consisting of open-ended questions about participants' experiences during the network penetration test, with participants in the informed condition explicitly asked about "the nature of deception on the network, if found" (Appendix A.1). On Day 2 participants were asked about their experience across both days and to rate tools available to them and their prior knowledge (Appendix A.2). In addition, each day participants were given a Cyber Task Questionnaire (CTQ, Appendix A.3) in which they were asked to rate and explain the level of confusion, self-doubt, confidence, surprise, and frustration they felt during the cyber exercise, with the Day 2 version including a question about belief in the presence of deception on the network.

General Questions about the Individual. These questions were designed to measure general information about an individual such as their demographic information (Appendix A.4) and cyber security experience (Appendix A.5). These items are of particular interest because they may help diagnose whether given effects found in the data set are due to the experimental manipulation or a particular individual's background (even given random assignment to condition). They could also help explain the factors relevant to particular performance characteristics (e.g., initial moves of a participant with over twenty years of experience versus two years of experience). We also asked participants who experienced cyber deception to complete a questionnaire designed to assess their responses to deception in a network penetration context.

Cognitive Battery. General cognitive ability (i.e., I.Q.) is traditionally the best predictor of individual job performance across job categories and situations [16, 17]. Measurement of additional, specific cognitive abilities may provide additional predictive value in the context of particular jobs, reflecting the specific processing required in these domains. This includes circumstances in which initial selection on general cognitive ability already occurs as part of an employment screening process [18]. Furthermore, non-cognitive attributes (e.g., personality characteristics) may provide additional predictive power [19]. Therefore, the battery for this study includes a number of tasks and questionnaires that go beyond general cognitive ability in order to allow a more comprehensive understanding of the abilities and attributes that are thought to characterize red teamers or be predictive of performance in the domain of network penetration (e.g, [4, 20, 3]).

Cognitive tasks included the Shipley-2 [21] as a measure of overall cognitive ability, the Sandia Progressive Matrices (SPM; [22]) as a measure of fluid intelligence (i.e., those aspects of intelligence that allow for adaptive reasoning and problem solving), the Over-Claiming Questionnaire (OCQ; [23]) as a measure of ability to distinguish real from fictional items and decision-making confidence, the Operation Span (O-Span; [24]) task as a measure of working memory (i.e., ability to maintain information in memory and inhibit distractors), the Remote Associates Task (RAT; [25]) as a measure of convergent creative thinking (i.e., generating atypical links between concepts in order to generate a solution to a problem), and a set of insight and analytical problems to solve [26] in order to assess proficiency at generating incremental solutions (analytical problems) and at reframing problems and approaching them from different perspectives (insight problems). Personality assessments included the Big Five Inventory (BFI; [27]) as a measure of openness to experience, conscientiousness, extraversion, agreeableness, and neuroticism, the General Decision-Making Style Inventory (GDMSI; [28]) as an indicator of the way in which individuals approach and make decisions, the Indecisiveness Scale (IS; [29]) in order to determine if participants tend toward intuitive speeded decisions or toward gathering as much information as possible, and the Need for Cognition (NfC; [30]) as measure of individuals' tendencies to pursue and enjoy the process of thinking.

Participants were also given the Karolinska Sleep Diary (KSD; [31]) to assess sleep quality for the night prior to administration, as some participants were required to travel prior to participation and may have experienced sleep disturbances which could impact task performance.

Physiological. Physiological data was collected using Empatica E4 wrist-based devices. The Empatica E4 collects heart rate information (including heart rate variability via blood volume pulse), motion-based activity (accelerometer), peripheral skin temperature (infrared thermopile), and galvanic skin response (electrodermal activity sensor). Physiological signals like these have been used to characterize and predict cognitive and physical states in a variety of settings.

4. Implementation

The experimental design received approval from each of the authors' institutional ethics review boards. No personal identifying information (PII) was collected and all experimental data was anonymized with subject IDs. No cyber task performance or HSR information will be provided back to any participants' employers.

4.1. Participants

Prior to traveling to our site for the two-day study, participants were provided the statement of work which indicated that they would be participating in a series of capture the flag style events to measure the effectiveness of defensive software on a simulated network. Participants could request, in advance, software tools, reference information (e.g., technical documents), and other computer files they wished to be included for the event. This ensured participants would have access to preferred tools and that all participants had access to the same tools. They were aware their actions would be monitored during the task and that they would be asked to complete a series of reports and questionnaires. Finally, they were informed that they would be required to sign a nondisclosure agreement stating they would not reveal information about the task, the network vulnerabilities, and the defenses encountered (including to other participants).

Upon arriving to the study, participants were asked whether they would also like to be part of a human subjects research study (HSR) as part of the cyber exercise. Those who opted in provided physiological and cognitive data in addition to the network penetration task and task-specific questionnaires; they were offered a $25 Amazon gift card for their participation. Six participants did not volunteer for the HSR portion. Those who opted out wrote an extended red team report, such that participants spent the same amount of time in the study regardless of the decision they made.

4.2. Procedure

The study took place over two consecutive days with up to ten participants per session. Each participant was assigned to an individual work station in the same room but divided into private, cubicle-style spaces. A proctor was always present in the room to answer questions and ensure participants worked independently. We attempted to group similar conditions (e.g., informed) during sessions to minimize cross-contamination. Participants worked on the same network environment within a given day (e.g., morning and afternoon on Day 1) but a new variation of the environment across days (e.g., Day 1 vs. Day 2). They were given an attack laptop for the network penetration task and an additional laptop with internet connectivity for research and the cognitive battery. The items in the cognitive battery were either completed with printouts or using E-Prime 3.0 software.

5. Sample Population Characteristics

Data was collected on 132 professional red teamers, 126 of whom agreed to participate in the HSR portion of the study. This section summarizes their responses to questions in the individual measures relevant to characterizing the sample population. The vast majority of our participants were male with English as their primary language. Most were under 35 years old and had a bachelor's degree as their highest level of education. Responses on the experience questionnaire indicated that participants were fairly evenly split between Linux and Windows users, although some chose to write-in Mac or a combination of operating systems. Most tended to work in groups of two to three people for engagements that last one to two weeks. However, there was substantial variance in these responses. The participants indicated the highest level of their expertise and years of experience in cyber security, network reconnaissance, and generalized defense practice. This is the skill set most necessary for the cyber task presented in the Tularosa study.

6. Results

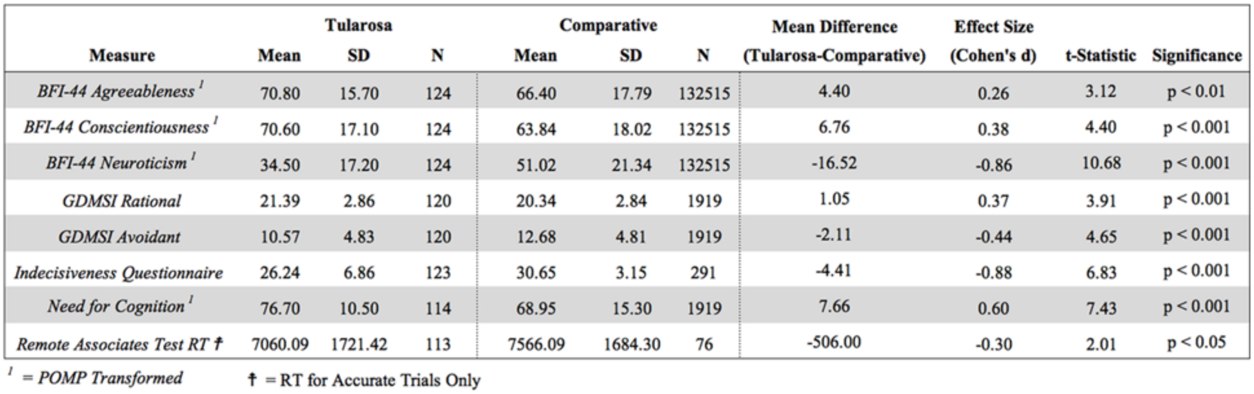
While the primary purpose of this paper is to describe the detailed methodology and data collected for the Tularosa experiment, we provide preliminary results addressing the cognitive battery, personality assessment, and cyber task questionnaire. These begin to address the first of the set of hypothesis around which the experiment was designed, as well as the validity of the experiment and the participant population.

6.1. Cognitive Battery/Personality Assessment

Following the cyber task on each day, participants completed a number of cognitive tasks and personality assessments. This battery was designed to characterize red teamers and to provide measurements of cognitive abilities and personality attributes previously hypothesized to be predictive of performance in this domain (e.g., [4, 20]). See Figure 2 for details of all significant effects from the comparative analyses.

Scores from the current work were compared against other data sets to determine how red teamers as a specialized population may differ from more general populations (e.g., college undergraduates). Means and standard deviations from our sample and comparison samples were calculated, as were mean difference scores, and an effect size (Cohen's d). Independent samples two-tailed t-tests were conducted to assess any statistical differences between groups. Assumptions of normality were not violated, however Welch's correction for unequal variances was applied since the sample sizes were often very different between groups.

For the GDMSI and NfC, our data was compared against that of 1,919 U.S. adults recruited via Amazon Mechanical Turk to assess attitudes toward privacy and security in the cyber domain [20]. Results suggest that for the rational subscale of the GDMSI, our sample displays a more rational decision making style relative to the comparison sample. For the avoidant subscale of the GDMSI, our sample shows a less avoidant style. These results indicate that network penetration professionals approach decision-making scenarios with a relatively high emphasis on a thorough search for and evaluation of alternative approaches while avoiding postponement of decision execution. An analytical



	Tularosa			Comparative			Mean Difference	Effect Size		
Measure	Mean	SD	N	Mean	SD	N	(Tularosa-Comparative)	(Cohen's d)	t-Statistic	Significance
BFI-44 Agreeableness [1]	70.80	15.70	124	66.40	17.79	132515	4.40	0.26	3.12	p < 0.01
BFI-44 Conscientiousness [1]	70.60	17.10	124	63.84	18.02	132515	6.76	0.38	4.40	p < 0.001
BFI-44 Neuroticism [1]	34.50	17.20	124	51.02	21.34	132515	-16.52	-0.86	10.68	p < 0.001
GDMSI Rational	21.39	2.86	120	20.34	2.84	1919	1.05	0.37	3.91	p < 0.001
GDMSI Avoidant	10.57	4.83	120	12.68	4.81	1919	-2.11	-0.44	4.65	p < 0.001
Indecisiveness Questionnaire	26.24	6.86	123	30.65	3.15	291	-4.41	-0.88	6.83	p < 0.001
Need for Cognition [1]	76.70	10.50	114	68.95	15.30	1919	7.66	0.60	7.43	p < 0.001
Remote Associates Test RT †	7060.09	1721.42	113	7566.09	1684.30	76	-506.00	-0.30	2.01	p < 0.05

[1] = *POMP Transformed* † = RT for Accurate Trials Only

Figure 2. Comparative Analyses: Statistics for responses on cognitive tasks and questionnaires.

and decisive approach has been suggested in prior characterizations of this group [4]. Results for the NfC scale suggest that our sample exhibits a higher need for cognition than the comparison sample, indicating that red teamers have a greater tendency than the comparison sample to pursue difficult problems and to enjoy the process of thinking, which is consistent with what prior interviews have implied [3].

For the IS, our participant results were compared against those from 291 undergraduate students [29]. Results suggest that our sample is less indecisive. These findings support the GDMSI result of a less avoidant style of decision making and are consistent with the notion that network penetration professionals tend to be decisive when presented with decision situations.

Our BFI was compared against a dataset [32] of 132,515 internet users living in the United States and Canada, aged 21-60. The sample of network penetration professionals exhibited higher scores on Agreeableness (predilection toward trust and compliance), Conscientiousness (level of efficiency and organization), and and lower scores on Neuroticism (an irritable, unhappy disposition), relative to the comparison dataset.

Our scores on the SPM were compared against a sample of 171 undergraduate students for the Day 1 session and a subset of 160 undergraduate students for the Day 2 session [33], matching groups who answered the same subset of problems. Our scores on the O-Span were compared against that of 6,236 college students [34]. No significant effects were observed in either case.

For the Insight and Analytical Problem Solving task and the RAT, the average solution rates (and reaction time for the RAT) were calculated based on the data of college students (see [26] for the Problem Solving Task; see [35] for the RAT) for the subset of problems we used. Standard deviations were not reported for the solution rates, so the only statistical comparison that could be attempted was on the RAT reaction time to produce correct solutions, which yielded a small effect where the Tularosa sample was faster.

6.2. Cyber Task Questionnaire

The primary research question that we begin to address with the Cyber Task Questionnaire (CTQ) involves the participants belief in deception given the manipulated two independent variables: Presence of cyber deception (absent vs. present) and Information about deception (uninformed vs. informed).

Responses to the belief in deception item (Appendix A.3) were coded using a scale from 1 = definitely no deception to 5 = definitely deception. Two raters completed the scoring, and scores were averaged across raters for analysis. Inter-rater reliability showed satisfactory reliability for Day 1 ratings (83% agreement, Cohen's κ = .77). At the end of each day, participants also reported the extent to which they felt confused, self-doubt, confident, surprised, and frustrated with the task on scales from 1 to 5.

Between-Group Differences on Day 1. We consider between-group differences on Day 1 to answer the research questions introduced in Figure 1 as *Planned Comparisons Between Groups*. A 2 (Cyber Deception Presence: absent vs. present) x 2 (Information: uninformed vs. informed) between-subjects ANOVA showed that there was a main effect for Presence, $F(1, 61) = 12.36$, $p < .001$, where those in the Present conditions reported more significantly belief in deception compared ($M = 3.60$) to Absent ($M = 2.19$), $p < .001$. There was also a nonsignificant trend for information, such that those informed about deception tended to suspect more deception $(M = 3.21)$ compared to those uninformed $(M = 2.58)$, $p = .125$. See Figure 3 for comparisons between each of the four experimental conditions.

Correlations between variables on Day 1 also provide interesting indications of different cognitive effects across conditions that mirror examples documented in the pilot studies. The non-deceptive conditions showed significant positive correlations between reporting frustration and confusion ($r = .574$, $p < 0.001$ for Control (A1) and $r = .454$, $p < 0.05$ for Cyber Aware (D1)). This could indicate that the task

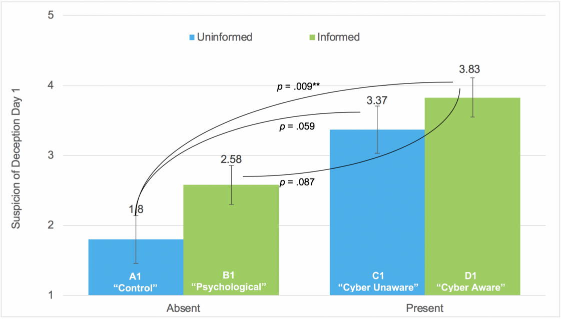


Figure 3. Day 1 between-group differences in belief of deception. Results suggest a cumulative effect of information and presence, such that veridical information in the presence of cyber deception instills the greatest belief in the presence of deception.

itself had confusing aspects which led to frustration among participants. Both Informed conditions showed significant positive correlations between reporting frustration and self-doubt ($r = .391$, $p < 0.05$ for present and $r = .583$, $p < 0.01$ for absent), and negative correlations between self-doubt and confidence ($r = -.536$, $p < 0.01$ and $r = -.511$, $p < 0.01$ respectively). This may indicate that information of the presence of deception (regardless of the veracity of the statement) can cause self-doubt to the participant, which affects confidence. Both Present conditions showed significant positive correlations between reporting frustration and surprise ($r = .563$, $p < 0.01$ for uninformed and $r = .708$, $p < 0.001$ for informed), as well as self-doubt and confusion ($r = .535$, $p < 0.01$ for uninformed and $r = .381$, $p < 0.05$ for informed), indicating that cyber deception may cause a cyber attacker surprise, and confusion about the network may lead to increased self-doubt when attacking. In the Cyber Unaware condition (C1), suspicion of deception was negatively correlated with self-doubt ($r = -.535$, $p < 0.05$), mirroring what was discovered in the pilot studies [11]; participants could be attributing task performance to feelings of inadequacy instead of to deception deployed on the network. However, in the Cyber Aware condition (D1), a positive correlation was observed with confidence ($r = .490$, $p < 0.05$), which could be reflecting the fact that since they were informed, and likely found deception evident on the network, they felt confident in their ability to negotiate it. In fact, an opposite, though marginal, effect was observed in the absent-informed condition, suggesting that being informed about deception but not finding anything on the network to support that claim resulted in less confidence about the attack strategy.

The CTQ data is a small portion of the data available for analysis. These answers were collected at the end of each day and required participants to think back and remember how they felt during the task and are only as reliable as the participants' memory and self-judgment. Self-reported thoughts and feelings often produce small effects, requiring greater sample sizes to detect significant differences between groups.

Next steps for this analysis would be to supplement these findings with participants' qualitative reports on their thoughts, feelings, and strategies used. Analyzing the daily briefings and Mattermost text will enhance the current data by providing a more coherent picture about how subjects experienced deception (and lack thereof). In addition, the emotional measures included in this data (e.g., frustration, surprise) should be compared to subjects' physiological responses. Finally, this data can be used to identify participants who felt highly confident in their performance, which will allow us to compare their actual success (as described in Section 8) with their perceived success.

Overall, employing real deception may evoke more frustration and suspicion regardless of whether attackers are told about the deception or not. Additionally, these data suggest that merely telling attackers that there may be deception can cause them to be more suspicious and more surprised about the network, possibly slowing them down or motivating them to change their strategy. Notably, the largest between-group disparity in many calculations was between the Control (A1) and the Cyber Aware (D1). This is particularly noteworthy because it may begin to counter popular opinion that cyber deception techniques are only/more effective if the attacker is unaware of its presence.

7. Discussion

While there are cyber games and Capture the Flag (CTF) activities that occur every year, we believe this is the largest controlled experiment which held constant the tools and exploits available to the attacker, the network topology and vulnerabilities, and the time participants had to launch attacks. In addition, the amount and variation of data collected is unprecedented (74 GB of data, plus an additional 1537 GB of screen capture video).

7.1. Design Decisions

Ferguson-Walter, Lafon, and Shade began to examine the effectiveness of decoys used for cyber defense with the pilot studies described in [11]. The results of those studies indicated that cyber deception had a measurable impact on attacker performance, with more time spent on decoys than real machines and self-reported confusion of which were the decoys. They also investigated whether just the belief that deception is in use can negatively affect attacks. We built upon the results of those pilot studies. Many of the aspects of the experiment were kept the same, but key changes were made to ensure a more rigorous experimental design. Other aspects were changed due to necessity rather than a focused improvement to the methodology; we discuss some trade-offs below.

Notably there were some calculated trade-offs in the design decisions for the Tularosa study that are less realistic than the pilot studies, the first being that the red teamers were asked to work independently. We opted for a larger sample size, trading off a degree of realism for those used to working as part of a team. For some participants, this may have hampered their performance; however, it was a restriction that was distributed across all participants and could even mimic a team that trades off roles across different tasks or days.

While originally framed as a CTF event, we decided not to include literal flags for the participants to collect along the way. We felt this could produce the wrong kind of motivation and could potentially be used as a "tell" that could disclose the deception in an unintended way. We investigated the use of "false flags" to be used as part of the deception but had concerns about biasing the results of the study by unfairly rewarding participants for interacting with the decoys. We did not use flags, believing this increased realism and reduced the chances of incorporating extra bias in the results. However, the lack of flags makes the data coding more complicated (as described in Section 8). It also caused some confusion amongst participants who had no clear metric to know when they were done with the cyber task. However, with only one day provided for each network, there was no concern that the participants would run out of things to investigate.

The next trade-off stems from having a large sample size in this study. While the Tularosa study generally collected vastly more data, the pilot studies did have trained observers in the room, encouraging participants to think aloud during the task. This enabled audio recording (and eventually text transcription with easy access to time stamps) to correlate verbal comments expressing frustration or confusion back to the computer interaction at that time. This gave the research team the ability to ask for clarification and also directly link the commands attempted and interactions with certain machines with psychological attributes of the participant. While this was a major strength of the pilot studies, it simply was not scalable. Additionally, with each participant working individually, we deliberately tried to minimize any verbal discussion or questions during the study that could influence another individual's thoughts or actions. We had a proctor in the room at all times, noting any blatant signs of frustration and noting the time. Additionally we had a Red Team Report at the end of each day. In order to try to get similar information that can be more easily correlated with the time stamps of the participant interactions during the cyber task, we asked participants to use the chat client to document their plans and progress, in real-time, during the task. There was a wide variance in how people used this tool and how well people followed instructions in general.

A trade-off that has potential effects on the generalization of the results is that the cyber task ran on a fully simulated network designed just for this study. There were no real users. There was no unexplained messiness or policies not being followed unless designed by us. This is clearly not realistic and may reduce the measured effectiveness, especially of the psychological deception, where pilot studies suggest the effects of being informed of deception benefit from the natural messiness present in an enterprise network.

7.2. Experimental Validity and Limitations

Validity concerns, including internal, external, and construct validity were considered when designing the Tularosa study. Internal validity was supported by using the same proctors throughout the experiment who read from a standardized script for instructions and responses to participant questions. Participant time on task was cataloged and monitored throughout the session, as were breaks, and lunch. Duration of the cyber task was controlled, as was the lunch break, for all participants. Participants were not allowed to discuss the cyber task during breaks. Identical copies of the cyber range were presented to all participants for a given condition, and identical machines (Kali Linux and internet-ready reference laptops). We arranged ahead of time to include any publicly available tools requested by participants, however no proprietary or costly tools were allowed. Additionally, a large standard set of red teaming tools were provided. A within subjects' component was implemented, whereby only cyber range deception was manipulated on Day 2. This design choice reduced the amount of individual variability across days and conditions inherent in between subjects' designs.

Many aspects of the Tularosa design support external validity. Since this was a tightly controlled laboratory study, the ecological validity could be called into question. For example, the standard set of tools provided could have hampered the performance of participants who were out of their comfort zones and unable to rely on tools they regularly use. As a proxy for one aspect of ecological validity, we asked participants to rate on a scale of 1-5 how they felt regarding the tools provided to them during the experiment (Appendix A.2). The mean rating was 3.51 out of 5, with a standard deviation of 0.93, suggesting that participants were largely satisfied with the tool selection provided to them. Participants were provided with a popular red teaming platform, Kali Linux, as well as internet access on a separate laptop for research. This experiment was designed to test the behavior of red teamers, and how this study would generalize to other populations who perform cyber attacks is unknown at this time. We subcontracted participants through various companies in several states around the United States, thus giving this project a broad, random sample within the specific population of professional red teamers. That said, this experiment was not an "in the wild" red teaming exercise, and thus proprietary tools were not allowed, participants had to work alone rather than in groups, and had a tightly controlled schedule. Finally, real-world cyber attack scenarios and ted teaming engagements typically exceed one day. Moreover, often the attacker will be the deciding factor of how long the engagement continues, which could change dynamically based on many relevant factors including interest, difficulty, and priority. We only allowed the participants to perform the task for one day per network. This was a monetary necessity but does diverge from the usual experience as evident in the data collected from the participants on the usual duration of engagements.

Construct validity is difficult to measure currently, as many planned future analyses will be required to determine if the deception led to altered cyber-behavioral performance. However, results discussed in Section 6.2 on self-reported suspicion of deception by condition did reveal associations between the cyber deception manipulation and suspicion. The data suggest an aggregate effect of the two deception manipulations, as the Cyber Aware condition showed the largest suspicion scores, whereas the Cyber Unaware condition produced an effect of roughly 80 percent that of the Cyber Aware condition. These data need to be scrutinized more carefully to disentangle the specific contributions of each of the deception manipulations.

8. Future Work

This paper described the experimental methodology and data collected in the Tularosa experiment. We presented some preliminary results, but many hypotheses posed still remain. These results will be presented in future publications. Future analysis will address the following hypotheses:

- Defensive (cyber and psychological) deception tools impede attackers who seek to penetrate computer systems and infiltrate information.
- Defensive cyber deception tools are effective even if an attacker is aware of their use.
- Defensive deception is effective if attackers believe it may be in use, even when it is not.
- Defensive (cyber and psychological) deception causes an attacker increased cognitive load.
- There is an observable correlation between cyber deception and physiological measures.

Performance measurement is key to addressing these hypotheses and there can be multiple (sometimes competing) indicators of success or failure for the cyber task. Future work includes evaluating vulnerabilities correctly identified and exploited (assessing both quantity and speed), as well as any mismatch between reported/perceived success and measured success. We will also calculate time until detection and time spent attacking each host (to assess time wasted on decoys).

In addition to the preliminary results presented in this paper, we will evaluate the cognitive and physiological data to identify signs of increased stress, cognitive load, or changes in emotional state. Ultimately, we will investigate whether increased confusion and frustration is correlated with lower performance in a network penetration task.

9. Conclusions

Cyber deception has been described as a "game changer" in cyber security–one that can allow the cyber defender to leverage the "home-field advantage" of owning and controlling the targeted network environment. The Tularosa study was designed to empirically measure the effectiveness of cyber (and psychological) deception on an attacker's ability to perform reconnaissance and exploitation. While this initial paper describes the experimental design, methodology, cyber range, participant population, and data collected, it only begins to address a subset of the research questions which motivated the work. Future publications will provide results addressing these main hypotheses.

The scale and breadth of data collected in this controlled study will allow for further future analyses beyond those described in this paper. Furthermore, there are many cyber defense research questions beyond the effectiveness of cyber deception that this data can help address.

Acknowledgments

We would like to thank all the professionals who participated in this study. We would like to thank members of the technical team: Rick van Tassel, Emmanuel Orozco, Kim Ta, Danielle Sanchez, Wellington Lee, SeanMicheal Galvin, Samuel Mulder, Armida Carbajal and Maxine Major. Dr. Dirk van Bruggen and Dr. Sunny Fugate went above and beyond to help make this study possible. Dr. Dana LaFon and Dr. Kristin Heckman gave early input on the design. Mary Berlage, Steve Danko and Aaron Pennington helped with project management. Dr. Robert Gutzwiller provided valued suggestions on the paper and its figures.

Opinions expressed are those of the authors and not necessarily those of the U.S. Government.

References

[1] K. E. Heckman, F. J. Stech, R. K. Thomas, B. Schmoker, and A. W. Tsow, *Cyber Denial, Deception and Counter Deception: A Framework for Supporting Active Cyber Defense*. Advances in Information Security, Springer International Publishing, 2015.

[2] N. Rowe and J. Rrushi, *Introduction to Cyberdeception*. Springer International Publishing, 2016.

[3] T. C. Summers, K. J. Lyytinen, T. Lingham, and E. A. Pierce, "How hackers think: A study of cybersecurity experts and their mental models," 2013.

[4] S. G. Campbell, P. O'Rourke, and M. F. Bunting, "Identifying Dimensions of Cyber Aptitude: The Design of the Cyber Aptitude and Talent Assessment," *Human Factors and Ergonomics Society Annual Meeting*, vol. 59, pp. 721–725, Sept. 2015.

[5] N. Provos, "A Virtual Honeypot Framework," in *USENIX Security Symposium*, 2004.

[6] S. T. Trassare, R. Beverly, and D. Alderson, "A Technique for Network Topology Deception," in *IEEE Military Communications Conference*, Nov. 2013.

[7] F. Cohen and D. Koike, "Feature: Leading Attackers Through Attack Graphs with Deceptions," *Computers & Security*, vol. 22, pp. 402–411, July 2003.

[8] P. Aggarwal, V. Dutt, and C. Gonzalez, "Cyber-Security: Role of Deception in Cyber-Attack Detection," in *Nicholson D. (eds) Advances in Human Factors in Cybersecurity. Advances in Intelligent Systems and Computing*, vol. 501, pp. 85–96, Springer, 2016.

[9] N. C. Rowe, E. J. Custy, and B. T. Duong, "Defending cyberspace with fake honeypots," *Journal of Computers*, vol. 2, pp. 25–36, Apr. 2007.

[10] M. L Bringer, C. Chelmecki, and H. Fujinoki, "A Survey: Recent Advances and Future Trends in Honeypot Research," *International Journal of Computer Network and Information Security*, vol. 4, Sept. 2012.

[11] K. J. Ferguson-Walter, D. S. LaFon, and T. B. Shade, "Friend or Faux: Deception for Cyber Defense," *Journal of Information Warfare*, vol. 16, no. 2, pp. 28–42, 2017.

[12] J. B. Michael, N. C. Rowe, H. S. Rothstein, T. C. Wingfield, M. Auguston, and D. Drusinsky, "Phase II report on intelligent software decoys: intelligent software decoy tools for cyber counterintelligence and security countermeasures," 2004.

[13] G. Wagener, R. State, A. Dulaunoy, and T. Engel, "Self Adaptive High Interaction Honeypots Driven by Game Theory," in *International Symposium on Stabilization, Safety, and Security of Distributed Systems*, pp. 741–755, Springer-Verlag, 2009.

[14] V. Nicomette, M. Kaâniche, E. Alata, and M. Herrb, "Set-up and deployment of a high-interaction honeypot: experiment and lessons learned," *Journal in Computer Virology*, vol. 7, pp. 143–157, May 2011.

[15] F. Cohen, I. Marin, J. Sappington, C. Stewart, and E. Thomas, "Red teaming experiments with deception technologies," *IA Newsletter*, 2001.

[16] M. J. Ree and J. A. Earles, "Intelligence Is the Best Predictor of Job Performance," *Current Directions in Psychological Science*, vol. 1, pp. 86–89, June 1992.

[17] F. L. Schmidt, "The Role of General Cognitive Ability and Job Performance: Why There Cannot Be a Debate," *Human Performance*, vol. 15, pp. 187–210, Apr. 2002.

[18] D. Lubinski, "Scientific and Social Significance of Assessing Individual Differences: "Sinking Shafts at a Few Critical Points"," *Annual Review of Psychology*, vol. 51, no. 1, pp. 405–444, 2000.

[19] F. L. Schmidt and J. E. Hunter, "The validity and utility of selection methods in personnel psychology: Practical and theoretical implications of 85 years of research findings," *Psychological Bulletin*, pp. 262–274, 1998.

[20] S. Egelman and E. Peer, "Predicting privacy and security attitudes," *ACM SIGCAS Computers and Society*, vol. 45, no. 1, pp. 22–28, 2015.

[21] W. C. Shipley, C. P. Gruber, T. A. Martin, and A. M. Klein, *Shipley-2 Manual*, vol. 30. Los Angeles, CA: Western Psychological Services, 2009.

[22] L. E. Matzen, Z. O. Benz, K. R. Dixon, J. Posey, J. K. Kroger, and A. E. Speed, "Recreating Raven's: software for systematically generating large numbers of Raven-like matrix problems with normed properties," *Behavior Research Methods*, vol. 42, pp. 525–541, May 2010.

[23] D. Paulhus, P. Harms, M. Nadine Bruce, and D. C Lysy, "The Over-Claiming Technique: Measuring Self-Enhancement Independent of Ability," *Journal of personality and social psychology*, vol. 84, pp. 890–904, May 2003.

[24] N. Unsworth, R. P. Heitz, J. C. Schrock, and R. W. Engle, "An automated version of the operation span task," *Behavior Research Methods*, vol. 37, pp. 498–505, Aug. 2005.

[25] A. Cropley, "In Praise of Convergent Thinking," *Creativity Research Journal*, vol. 18, pp. 391–404, 2006.

[26] M. Wieth and B. D. Burns, "Incentives improve performance on both incremental and insight problem solving," *Quarterly Journal of Experimental Psychology*, vol. 59, pp. 1378–1394, Aug. 2006.

[27] O. P. John and S. Srivastava, "The Big-Five trait taxonomy: History, measurement, and theoretical perspectives.," in *Handbook of Personality: Theory and Research*, vol. 2, pp. 102–138, New York, NY, USA: Guilford Press, l. a. pervin & o. p. john ed., 1999.

[28] S. G. Scott and R. A. Bruce, "Decision-Making Style: The Development and Assessment of a New Measure," *Educational and Psychological Measurement*, vol. 55, pp. 818–831, Oct. 1995.

[29] E. Rassin, P. Muris, I. Franken, M. Smit, and M. Wong, "Measuring General Indecisiveness," *Journal of Psychopathology and Behavioral Assessment*, vol. 29, pp. 60–67, Mar. 2007.

[30] J. T. Cacioppo, R. E. Petty, and C. F. Kao, "The Efficient Assessment of Need for Cognition," *Journal of Personality Assessment*, vol. 48, pp. 306–307, June 1984.

[31] T. Åkerstedt, K. Hume, D. Minors, and J. Waterhouse, "The Subjective Meaning of Good Sleep, An Intraindividual Approach Using the Karolinska Sleep Diary," *Perceptual and Motor Skills*, vol. 79, pp. 287–296, Aug. 1994.

[32] S. Srivastava, O. P. John, S. D. Gosling, and J. Potter, "Development of personality in early and middle adulthood: Set like plaster or persistent change?," *Journal of Personality and Social Psychology*, vol. 84, no. 5, pp. 1041–1053, 2003.

[33] V. Clark, "Darpa ram replay." private communication, 2014. DARPA/ARO Contract W911NF-16-C-0018.

[34] T. S. Redick, J. M. Broadway, M. E. Meier, P. S. Kuriakose, N. Unsworth, M. J. Kane, and R. W. Engle, "Measuring Working Memory Capacity With Automated Complex Span Tasks," *European Journal of Psychological Assessment*, vol. 28, no. 3, pp. 164–171, 2012.

[35] E. M. Bowden and M. Jung-Beeman, "Normative data for 144 compound remote associate problems," *Behavior Research Methods, Instruments, & Computers*, vol. 35, pp. 634–639, Nov. 2003.