CMU-IT-18-227

ACQUISITION RESEARCH PROGRAM SPONSORED REPORT SERIES

Risk Management and Information Assurance Decision Support

4 March 2019

Hanan Hibshi

Dr. Travis Breaux

Institute for Software Research
Carnegie Mellon University

Disclaimer: This material is based upon work supported by the Naval Postgraduate School Acquisition Research Program under Grant No. N00244-17-1-0012. The views expressed in written materials or publications, and/or made by speakers, moderators, and presenters, do not necessarily reflect the official policies of the Naval Postgraduate School nor does mention of trade names, commercial practices, or organizations imply endorsement by the U.S. Government.

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

Acquisition Research Program
Graduate School of Business & Public Policy
Naval Postgraduate School

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

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

Acquisition Research Program
Graduate School of Business & Public Policy
Naval Postgraduate School

Abstract

Department of Defense (DoD) information assurance (IA) certification and accreditation relies on a multi-tier risk framework where security assessment aligns with NIST information assurance control set. The human analyst faces multiple burdens, including resolving dependencies among IA controls, understanding how security requirements apply to a specific context, and integrating expertise from multiple technical areas. In this research, we will investigate new ways to leverage component-based architecture in reducing security threats. These new techniques integrate human security expert judgements with notions of composable security to identify interactions among security requirements that affect overall system assurance levels. The research is based on using the Multi-factor Quality Measurement (MQM) method to collect security ratings from multiple experts with documented expertise in specific technical areas. We will share results from collecting and analyzing data from security experts with an average of 10 years of experience. The results of this evaluation will improve DoD acquisition by providing reliable ways to express and evaluate cybersecurity mitigations that are commensurate with changing security risks. These evaluations will be semi- automated, focusing expert evaluations on relevant details in an IT scenario. In addition, the MQM framework can be extended and reused for security, privacy, and even outside of security for domain where composable requirements exist.

This research will yield important public benefits to private sector companies who supply and consume the dual-purpose information technology (IT) used by the DoD and who are frequently subject to security threats from organized crime, foreign governments and stateless hackers. This work helps IT by providing companies new means for security risk assessment that collects multiple experts input in a feasible approach without the hassle of hiring more experts. The ratings provided by experts can support companies with their security risk assessment and related security decisions. The experts can also provide further suggestions through the tool which can help companies identify unforeseen dependencies and/or missing requirements.

THIS PAGE INTENTIONALLY LEFT BLANK

About the Authors

Hanan Hibshi - is a Research and Teaching Scientist at the Information Networking Institute at Carnegie Mellon University. Dr. Hibshi's research area includes: security requirements, usable security and expert's decision-making. Dr. Hibshi's research involves using grounded theory and mixed-methods user experiments to extract rules for use in intelligent systems. Dr. Hibshi received a PhD in Societal Computing from Carnegie Mellon University, an MS in Information Security Technology and Management from the Information Networking Institute at Carnegie Mellon University, and a BS in Computer Science from King Abdul-Aziz University in Jeddah, Saudi Arabia. [hhibshi@cmu.edu]

Travis Breaux - is an Associate Professor of Computer Science, appointed in the Institute for Software Research of the School of Computer Science at Carnegie Mellon University. Dr. Breaux's research program searches for new methods and tools for developing correct software specifications and ensuring that software systems conform to those specifications in a transparent, reliable and trustworthy manner. This includes demonstrating compliance with U.S. and international accessibility, privacy and security laws, policies and standards. Dr. Breaux is the Director of the Requirements Engineering Laboratory at Carnegie Mellon University. Dr. Breaux has several publications in ACM and IEEE-sponsored journals and conference proceedings. Dr. Breaux is a member of the ACM SIGSOFT, IEEE Computer Society and USACM Public Policy Committee. [breaux@cs.cmu.edu].

THIS PAGE INTENTIONALLY LEFT BLANK

CMU-IT-18-227

ACQUISITION RESEARCH PROGRAM SPONSORED REPORT SERIES

Risk Management and Information Assurance Decision Support

4 March 2019

Hanan Hibshi

Dr. Travis Breaux

Institute for Software Research
Carnegie Mellon University

Disclaimer: This material is based upon work supported by the Naval Postgraduate School Acquisition Research Program under Grant No. N00244-17-1-0012. The views expressed in written materials or publications, and/or made by speakers, moderators, and presenters, do not necessarily reflect the official policies of the Naval Postgraduate School nor does mention of trade names, commercial practices, or organizations imply endorsement by the U.S. Government.

THIS PAGE LEFT INTENTIONALLY BLANK

Table of Contents

THIS PAGE LEFT INTENTIONALLY BLANK

Introduction

Despite the abundance of well-documented guidelines and checklists, such as the National Institute Standards and Technology's (NIST) 800 special publication series ("NIST/ITL Special Publication (800)," 2015), the process still relies on human security analysts. Security guidelines in the form of checklists are an important part of compliance and audit processes including the acquisition process at the US Department of Defense (DoD). The analysts, who reason about security and assess the risk, do not treat security controls as independent factors, they, instead, reason over potentially millions of scenarios that account for various permutations of network type, services offered, threat type, etc. When mapping the security requirements in the checklist to scenarios and components found in real systems, security analysts do not treat requirements independently, they account for the priorities and dependencies that exist between the security requirements. This effect of context and requirements composition on security requirements expert ratings has previously been examined by Hibshi et al. (Hibshi, Breaux, & Broomell, 2015).

The reliance on human experts to comprehensively assess the security of systems becomes a burden with the global shortage of security experts. In 2018, NIST's report to the President of the United States titled: "Supporting the Growth and Sustainment of the Nation's Cybersecurity Workforce" states that the global shortage in cybersecurity workforce is projected at 1.8 million by 2022 (National Institute Standards and Technology, 2018). The scarcity of experts and the need for cybersecurity as the number of information security incidents keeps increasing, makes the provision of intelligent decision support and semi-automated solutions a necessity.

It is also important to note that security experts are diverse in terms of their background knowledge, because security expertise crosses different domains, such as hardware, software, cryptography, and operating systems. Hibshi et al. asked security experts to analyze security artifacts and compared patterns of situation awareness between experts and novices. The experts were able to demonstrate better decision

making compared to novices by exhibiting more confident patterns of situation awareness (Hibshi, Breaux, Riaz, & Williams, 2016).

The DOD acquisition process also relies on security analysts to review the controls. These analysts rely on their own expertise and background knowledge to reason over scenarios that account for various permutations of controls. For example, under NIST SP 800-53, the analyst decides if a specific system is high, medium or low impact and then the analyst satisfies the impact rating by selecting security controls (e.g., audit events, lock sessions, etc.). The DOD considers NIST security controls to be the minimum and requires additional sets of controls that vendors need to meet before they can work on classified networks(Swenson, 2009). Each control represents a class of technology aimed at mitigating a security threat.

As Hibshi et al. points out in their research, security assessment can be impacted by *context,* where security requirements apply*; priorities* of some requirements over other requirements; *uncertainty* due to human experts' memory constraints; and the *stove-piped knowledge* among security experts who come from a variety of backgrounds such as: systems, networks, databases and Web applications (Hibshi et al., 2015). In prior work, Hibshi et al. used factorial vignettes, wherein requirements and system constraints are variables in a scenario description. That work is limited, since the vignettes were only applied narrowly to website access, there was no guidance on how to select vignette variables, and new requirements were not evaluated for security impact. Moreover, the ratings were elicited from graduate students, and not security professionals.

We can summarize challenges that faces security risk assessment and decision-making as follows:

- The experts varying level of expertise and their stove-piped security knowledge and background.
- The composition of requirements corresponding to components of a system, and the varying priorities among requirements. Some requirements have higher priorities than others, depending on their strength in mitigating threats.
- The uncertainty in security decisions, that could result from ambiguity in abstract terminology that could lead to different experts interpreting the same requirement differently.
- The scarcity of security experts (U.S. Bureau of Labor Statistics., 2016).

In this technical report, we will explain how we build on previous research to address the above challenges by introducing the Multifactor Quality Measurement (MQM) method. The MQM, can be extended and reused for security, privacy, and even outside the security domain where composable requirements exist. An analyst or a researcher who aims to study a quality of interest can create scenarios and follow the steps defined in the MQM framework. By using MQM, one can examine the dependencies among requirements and collect additional missing requirements. In the upcoming sections of this report we will provide more details about the MQM framework. We will start with a literature review to discuss background and related work. Then, we will present our research methodology; followed by a description of the phases of the MQM. Next, we will present the research results, and threats to validity. We discuss our findings and we finally conclude with remarks highlighting future research directions.

THIS PAGE LEFT INTENTIONALLY BLANK

Background and Related Work

The MQM method discussed in this technical report combines research methods and knowledge from multiple disciplines: security, requirements engineering, statistics, and social science. In this section of the report we will review relevant literature that highlights the challenges in security risk assessment. Then, we present related work in scenario-based approaches in requirements engineering. Finally, we present background on factorial vignettes, which is a method well-known in social science that we adapt to generate security scenarios

Challenges in Security Risk Assessment

As explained earlier in our introduction, we define four challenges that faces security risk assessment: experts' stove-piped knowledge, requirements composition, the uncertainty in experts' decisions, and the scarcity of security experts. Below, we will discuss background and related work for each of these four challenges.

Security Analysts Expertise

Security problems are often assessed by experts who are responsible for reviewing a system specification, and deciding what mitigations will mitigate security threats. Experts are also responsible for making sure companies' security practices follow security guidelines, such as NIST 800-53. Security knowledge can be acquired from specialized courses, on-the-job training, or self-study. In addition, some experts may be more specialized in certain areas of security, such as web-security or mobile security. Ben-Asher and Gonzalez (Ben-Asher & Gonzalez, 2015) examined how the knowledge gap between novices and experts affect the analyst ability to detect cyber-attacks, as the experts performed significantly better than novices. To detect attacks successfully, cybersecurity experts need: 1) domain knowledge (Chi, 2006; Ericsson & Lehmann, 1996; Goodall, Lutters, & Komlodi, 2009) that is obtained through formal academic learning and practical hands-on experience with tools; and 2) situated knowledge which is organization dependent and which analysts tend to learn through continuous interaction with certain

environments (Goodall, Lutters, & Komlodi, 2004; Goodall et al., 2009; Schmidt & Hunter, 1993).

Security Requirements Composition

Understanding complex security attacks requires knowledge combined from a number of security fields to help analyze how the ``pieces of the puzzle'' compose together (Ben-Asher & Gonzalez, 2015; Hibshi et al., 2015). Stuxnet (Chen & Abu-Nimeh, 2011)is a good example where the attack targeted networks with hosts that run the Windows operating system and Siemens Step7 software. This attack, which targets vulnerabilities found on network hosts, proves that focusing on strengthening the security of the network alone is not sufficient as other factors, such as the hosts, their operating systems, and other connected components, need to be taken into consideration when performing the security risk assessment (Garfinkel, 2012). This broad understanding helps analysts to determine the proper requirements that work together to mitigate attacks. For example, stronger passwords with rules of 16 alphanumeric and special characters could be considered a good security requirement, but this cannot be an absolute rule. The type of password relies on other factors such as: the type of network where the connection is made, the sensitivity of the data involved, and so on (Hibshi et al., 2015).

Uncertainty in Security Decision-Making

The research paradigm in software engineering is shifting towards recognizing uncertainty as a first-class concern that affects design, implementation, and deployment of systems (Garlan, 2010). Garlan argues that the human in the loop, mobility, rapid evolution, and cyber physical systems are possible sources of uncertainty (Garlan, 2010). These sources of uncertainty affect the analyst security assessment. Michels et al. proposed a probabilistic first-order logic model to provide digital support tools for human operators when reasoning about objects given uncertain information (Michels, Velikova, Hommersom, & Lucas, 2013). The authors applied their proposed model to simulated and real-time vessel data and their results helped in reasoning about uncertainty originating from missing information in the dataset (Michels et al., 2013). It remains

unclear, how the proposed model would handle uncertainty originating from the decisions made by the human operators.

We focus on the human uncertainty in expert security assessments that could be interpersonal and intrapersonal. Interpersonal uncertainty exists between different experts as experts can judge the same situation differently. Intrapersonal uncertainty is the uncertainty within an analysts own judgment (Mendel, 2001). For example, an expert might describe a security requirement to be adequate. The uncertainty that this expert has about whether the combination of factors are themselves adequate is intrapersonal uncertainty, because the same experts might provide different judgments in two different times. The interpersonal uncertainty would be between two different experts would have different judgments of the situation and could disagree on the efficacy of the security requirement to mitigate an attack. Lipshitz and Strauss showed that uncertainty in decision making could be caused by inadequate understanding, incomplete information, or undifferentiated alternatives (Lipshitz & Strauss, 1997). They further argued that assumption-based reasoning and weighing pros and cons of competing alternatives are two possible strategies that decision-makers apply to coping with the uncertainties (Lipshitz & Strauss, 1997). In the work we present in this paper, security experts assess security in scenarios that compose multiple requirements. When faced with uncertainly in one requirement, for example, experts may analyze other requirements in the scenario to weigh the pros and cons of their decision.

Security Experts Scarcity

The number of security experts in the world is scarce. According to the U.S. Bureau of Labor statistics, there is around 100,000 information security analysts in the U.S. in 2016, earning a median income of $95,510 a year (U.S. Bureau of Labor Statistics., 2016). Employment is projected to grow by 28% by 2026, which is a faster growth rate then average (U.S. Bureau of Labor Statistics., 2016), and 56% growth in demand for security analysts is projected by 2026 (U.S. Bureau of Labor Statistics., 2016).

Scenario-based Approaches in Requirements Engineering

Scenario-based techniques have been argued to provide richer details needed in analyzing dependencies between system components and the environment when modeling human uncertainties (Sutcliffe, 1998), and in eliciting actual user needs and unforeseen requirements (Colin Potts, Takahashi, & Anton, 1994; Sutcliffe, 1998). Scenarios can either originate from the stakeholders' real practices before the system is designed (Sutcliffe, 1998), such as by pursuing the inquiry cycle model (Colin Potts et al., 1994); or they can originate from the system's specifications and design (Sutcliffe, 1998), such as use cases (Graham, 1996; Jacobson, 1992), misuse cases (McDermott & Fox, 1999) and Secure Tropos (Liu, Yu, & Mylopoulos, 2002; Mouratidis & Giorgini, 2007). Our factorial vignette-based approach uses scenarios to describe an environment that mimics reality to the security analyst to discover dependencies among requirements and elicit previously unforeseen requirements that mitigate threats. Unlike the inquiry cycle model that searches the requirements space in multiple directions by asking what, how, where, when and why questions (Colin Potts et al., 1994), our approach first collect answers to some of these questions and then shows these answers to expert. Then, the quality is measured to evaluate the strength of the answer toward affecting the overall system behavior, and this is obtained by experts' evaluations.

Scenario-based research in requirements engineering, including the work noted above, share a common feature: the ad-hoc starting point of scenario creation, which is generated by the analyst, is then re-inspected or refined. In MQM, a similar approach is adopted, which we call bootstrapping, wherein an analyst designs an initial scenario from their limited domain knowledge. Because the method guides the analyst toward increases in quality, this approach is preferable to an otherwise unbounded process with no clear guidance on where to stop generating more scenarios. Letier et al. proposed a scenario-based technique for requirements analysis and they indicate that adding both positive and negative scenarios results in large unstructured models (Letier, Kramer, Magee, & Uchitel, 2005).

Potts distinguishes abstractionism, where researchers rely on formal models; and contextualism, where the context of the system is well understood before deriving

requirements (C. Potts, 1997). Potts argues that abstractionist approaches use simplified models of a phenomena that leave out stakeholders' needs and requirements. Contextualist approaches use rich details resulting in creating systems that reflect the context of use and satisfy the stakeholder needs in the short term, but can be expensive and time consuming to progress or scale the design over time. Potts suggests that to build more useful systems, researchers should focus on approaches that integrate the two philosophies. The integration should adopt a strong committed view (C. Potts, 1997; C. Potts & Newstetter, 1997) in which, intangible phenomena are not explicit (Lincoln & Guba, 1985; C. Potts, 1997), but rather they are implicit in the stakeholders' interpretation of the domain (C. Potts, 1997; C. Potts & Newstetter, 1997). For example, business processes are not tangible, as they exist in the stakeholders' interpretation of a business (C. Potts, 1997). In our work, we acknowledge that security requirements in practice exist in the security analysts' interpretation of a system and their judgment of the situation, and that organizations rely on security experts' recommendations. The MQM method introduces structured scenario design using textual templates, which is different from strict contextual designs such as natural inquiries (Lincoln & Guba, 1985; C. Potts & Newstetter, 1997) that rely on rich descriptions, but could be time consuming.

Factorial Vignettes

The vignette experiments used in MQM are based on *factorial vignettes*, which are scenarios comprised of discrete factors that contribute to human judgment. Researchers systematically manipulate the factors to understand their composite and individual effects on a decision (Rossi & Nock, 1982; Wallander, 2009). Factorial vignettes are proven more effective to understanding decision-making than direct questioning or single statement ratings that obscure the underlying contributions of different factors to the overall decision (Alexander & Becker, 1978; Rossi & Nock, 1982; Wallander, 2009). In addition, the use of factorial vignettes, increases experimental realism as participants react to scenarios that are similar to what a participant may experience in the real world (Auspurg & Hinz, 2014).

Factorial vignettes are presented in surveys and user experiments using a basic template that contains multiple dimensions of the construct of interest. In our case, each

dimension is a security requirement that influences the perceived level of security risk: some requirements increase risk, while others decrease risk. We will show an example of the template used in our MQM study in the upcoming sections.

Research methods using factorial vignettes have been applied in social and decision science, psychology, sociology, and marketing, to name a few (Auspurg & Hinz, 2014; Wallander, 2009). We will highlight below related work that uses factorial vignettes as a research methodology.

McKelvie et al. used factorial vignettes to investigate the effect of different types of uncertainty on the decision-making of entrepreneurs in software industry (Auspurg & Hinz, 2014; McKelvie, Haynie, & Gustavsson, 2011). Based on their results, the authors argue that entrepreneurs prefer to avoid uncertainty, but the extent of that avoidance is affected by the type of uncertainty, the magnitude of the decision, and the domain expertise (Auspurg & Hinz, 2014; McKelvie et al., 2011). The authors argue that having their participants provide judgments to scenarios that consist of underlying factors, is an approach found to provide more accurate and less biased data when compared to other methods such as participant introspection (McKelvie et al., 2011). In our work, we are also interested in investigating uncertainty, risk, and expertise, but in the security domain. We find factorial vignettes an approach that allows participants to rate scenarios composed of multiple security configurations. The analysis of participant data will help investigate the composed requirements in the scenarios and explain their effect on security expert decisions.

Researchers have applied factorial vignettes to study different factors that impact cybersecurity. To study compliance with cybersecurity policy, researchers used factorial vignettes to explore factors that lead to policy violations (Johnston, Warkentin, McBride, & Carter, 2016; McBride, Carter, & Warkentin, 2012; Trinkle, Crossler, & Warkentin, 2014). Gomez and Villar use factorial vignettes to study the effect of uncertainty on dealing with cyberthreats (Gomez & Villar, 2018). Factorial vignettes have also been used to examine end-user security decision-making (e.g. file download) and explore their security risk perception (Hardee, West, & Mayhorn, 2006).

We have applied factorial vignettes to a new application domain: security requirements. We treat security requirements as factors and we manipulate these requirements by using specifications that should increase or decrease security in a vignette. Prior work mentioned above has focused on studying end-user related factors such as: effect of personality traits on employees compliance with company's information security policy, presence of compliance policies on the likelihood of playing online games (Trinkle et al., 2014), trust in cyberspace (Gomez & Villar, 2018), and effect of gain-and-loss on end-user security decisions (Hardee et al., 2006).

Except for the research conducted by Gomez and Villar (Gomez & Villar, 2018), prior work mentioned above that used factorial vignettes have focused on end-users. Gomez and Villar recruited computer science university students and treated them as the experts in their online experiments (Gomez & Villar, 2018). In our research, we focus on security experts. We will show in the upcoming sections of this report how we focus on recruiting industry experts. In addition to self-reported expertise questions, we include a security knowledge test in our studies to be able to assess a participant's security expertise.

THIS PAGE LEFT INTENTIONALLY BLANK

The Multifactor Quality Measurement

We now describe the Multifactor Quality Measurement (MQM) method for eliciting system constraints that affect overall quality. In prior work (Hibshi et al., 2015), we presented an empirical evaluation of using factorial vignettes for collecting security and found it to be effective. In this paper, we are integrating the technique into a framework, the MQM, that can be extended and reused outside of security. Figure 1 shows the different stages of the MQM. In addition, we address the limitations of prior work in the following way:

1. We evaluate the MQM across four domains: networking, operating systems, databases and web applications. In prior work (Hibshi et al., 2015), only one domain was evaluated (computer user surfing the web).
2. Participants are put in an expert role in the scenario (e.g. network administrator)
3. We recruit security experts from industry and government.

We now describe each phase of the MQM.

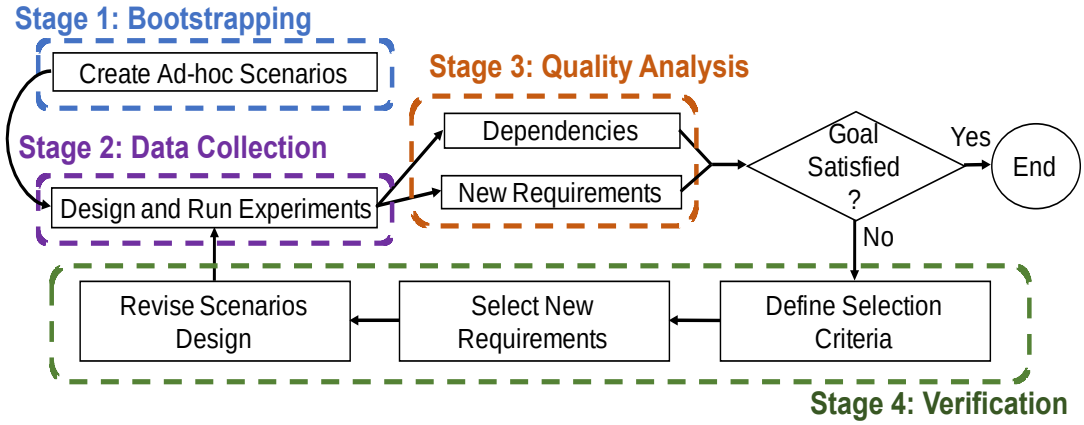

Figure 1: Multifactor Quality Measurement Method Overview

Stage 1: Bootstrapping

During bootstrapping, an analyst first chooses the quality to evaluate, and then the analyst chooses an initial scenario that describes a cohesive system viewpoint (Nuseibeh, Kramer, & Finkelstein, 1994). The ad hoc scenario is selected by the analyst who might

have limited knowledge, because the MQM will collect empirically measured improvements in this stage. This scenario is a text-based system description that includes the ways people interact with the system. We show an example scenario template in Figure 2.

Using scenarios in the MQM is based on a research method well-known and adapted in social science known as: *factorial vignettes* (Wallander, 2009). Factorial vignettes are scenarios comprised of discrete factors that contribute to human judgment. Researchers systematically manipulate the factors to understand their composite and individual effects on a decision (Rossi & Nock, 1982; Wallander, 2009). Factorial vignettes are proven more effective to understanding decision making than direct questioning or single statement ratings that obscure the underlying contributions of different factors to the overall decision (Alexander & Becker, 1978; Rossi & Nock, 1982; Wallander, 2009).

Figure 2 shows a template from the web applications security domain that consists of variables preceded by the (`$`) sign. A variable in the scenario is a security requirement category. The variables are replaced by different values that correspond to constraints on the system. The manipulation of variables and their values allows the analyst to generate different instantiations of the template, called vignettes, which will increase the number of scenarios that can be evaluated at one time. The `$WebAuth` variable represents the type of authentication used in the web application and it can take one of many values. To illustrate, we consider two extremely different values: “basic authentication,” which is a weak form of web-based authentication, or “form-based authentication using encrypted credentials stored in a database,” which is stronger. Similarly, the `$StoredUserData` variable represents how the user input is being collected, and could take the values: “collect user-supplied content from GET request,” or “require CSRF tokens and escape and validate user-supplied content from POST requests before storing;” and again, the latter value is stronger than the former.

```
You are a website administrator responsible for securing a web app against
cyberattacks. Currently, you are evaluating the following settings:
- The web app performs $WebAuth.
- The web app will $StoredUserData in a database for display to other users.
The Cross-Site Request Forgery attack is a serious security concern. Please answer
the following questions with regards to mitigating this threat.
```

Figure 2: Example Scenario Template from the Web Applications Domain

Study participants are asked to rate the adequacy of the overall security of the scenario on a 5-point scale where point 1 is labeled "inadequate", point 3 is labeled "adequate" and point 5 is labeled "excessive." This generates the `$Overall` dependent variable. The adequacy scale was evaluated by Hibshi and Breaux in a separate study (Hibshi & Breaux, 2016) and applied in prior work (Hibshi et al., 2015). Similarly, we ask users to provide ratings for the individual security requirements in the scenario, which generates a dependent variable for each rated requirement. For example, the web applications study has the `$WebAuthRating`, and the `$StoredUserDataRating`, which are the dependent variables representing experts' ratings of the `$WebAuth`, and `$StoredUserData`, respectively.

After creating the initial ad-hoc scenario, the analyst decides the number of factors and factor levels in the scenario:

Factors per domain: a domain could have its own subset of factors, with the possibility of having factors that are shared among different domains. The factors often correspond to categories of system constraint e.g., passwords, authentication type, etc. In addition, factors may, but do not necessarily have to, cross multiple domains, e.g., passwords affect databases, networks, and systems.

Levels per factor: how many levels will be manipulated. The levels, which correspond to technically specific interpretations of the factor, can be chosen as high or low levels. The goal is to choose levels that experts can distinguish to measure an effect or interaction among different levels. For example, if password complexity has high and low levels, we can measure whether password complexity affects overall security adequacy in conjunction with other security constraints.

Deciding on the number of factors depends on the quality of interest, the cost of running the surveys, and the estimated number of experts available to rate the scenarios against the quality of interest. An analyst would need to conduct a priori statistical power analysis to decide on the right number of factor/level combinations. Initial pilot studies and focus groups can also help with the design decisions in the bootstrapping phase as it would help eliminate unrealistic factor and level combinations (Hibshi et al., 2015).

We are not limited to one template, in addition to the web application template shown in Fig. 2, we use more templates and integrate factors and levels for three more security domains: Networking, systems and databases. All domains are shown in Table 1 along with the factors and levels. The general template that we use to generate different templates for different security domains is shown in Figure 3, below.

```
A popular online retailer offers a wide variety of products for purchase.  User
information in the company's databases includes consumers' credit card information
for purchasing products in the future.
 You are a $Domain administrator for the retailer who is responsible for securing
 the $Domain against cyberattacks. Currently, you are evaluating the following
 settings:
 - $Factor1
 - $Factor2 …
 The $Threat attack is a serious security concern. Please answer the following
 questions with regards to mitigating this threat.
```

Figure 3: Text template for the four security domains

Domain experts may suggest additional unforeseen requirements that would improve the measurements. An analyst could elicit new expert requirements from experts to improve the measurements. For example, security experts could provide more mitigation that would increase the adequacy ratings, so, we ask experts to list additional mitigations that they believe will increase security.

Stage 2: Data Collection

Once the scenarios are ready, the analyst finalizes the design of the overall experiment. This includes deciding which factors are between-subject or within-subject factors. The analyst in this stage decides on how to operationalize the survey: recruitment methods (e.g. in person, online, mailing lists), tools to be used, and whether expertise

screening questions are needed (e.g. knowledge tests, demographics). Finally, the analyst deploys the survey and starts data collection.

In our study, we recruited security experts who attended the SANSFIRE 2016 conference at Washington, DC. The SANS is a security research and education company that offers security training and certification to government and industry security analysts ("SANS Institute: About," 2017). We compensated each participant with a $25 Amazon gift card.

To better understand the expertise of our target population, security experts, we designed our surveys such that upon completion of the security ratings, participants are asked to take a security knowledge test (14 questions); and answer demographics questions (e.g. gender, age, experience, etc.)

TABLE I. USER STUDY SECURITY DOMAINS AND THEIR CORRESPONDING REQUIREMENT VARIABLES

Domain	Threat	Factor	Level Code	Level description
Network	Man-in-the-middle	**$NetworkAccess**	onsite	Onsite access using Ethernet
			offsite	Offsite external access through a secure VPN
		$NetworkAuth	simp6	A standard 6-digit password
			comp16	16-char password that must include an uppercase letter, lowercase letter, a symbol, and a number
			multi8	An 8-character alphanumerical password and a one-time password sent to a mobile phone
		$DMZ	allnosplit	DMZ contains the webserver, app server and the database server.
			split	DMZ contains the front-end webserver and the app server. The DB server is behind the firewall on the internal network. The app server communicates with the DB over a VPN.
Systems	malware	**$SocialMedia**	permit	Workstations permit access to social media sites
			prohibit	Workstations prohibit access to social media sites
		$AdminPriviledges	noauth	Prior to installing new software, employees who are local system administrators, are not required to re-authenticate
			auth	Prior to installing new software, employees are required to re-authenticate
		$VirusScanner	files	Workstations has programs to scan files against known malware signatures
			filesmem	Workstations has programs to scan memory and files against known malware signatures
			filesmempro	Workstations has programs to scan memory, files and processes against known malware signatures
Database	Privilage escalation	**$DBAccess**	extserver	User accounts and access control are handled by SQL table authentication
			sqlauth	User accounts and access control are handled by Windows Active Directory
		$DBMonitor	available	Database activities are logged
			needed	Database activities are logged, and inspected as needed (e.g., to examine a certain incident)
			month	Database activities are logged, and inspected each month by a trained auditor
		$Error	User	Errors are handled by notifying users who can then report the error message, as needed
			nouser	Errors are handled by logging the error message with no external notification to users
Web applications	Cross-site-request forgery	**$WebAuth**	basic	Basic authentication
			form	Form-based authentication using encrypted credentials stored in a database
		$StoredUserData	get	store user-supplied content from GET requests
			post	store user-supplied content from POST requests
			cpost	require CSRF tokens for user-supplied content from POST requests before storing
			cescpost	require CSRF tokens, escape and validate user-supplied content from POST requests before storing

Stage 3: Quality Analysis

In this stage, the analyst uses regression analysis to discover the weights of the factor levels (e.g., $WebAuth and the $StoredUserData) and to discover any interactions among the variables. The priorities of requirements are decided based on the weight of the coefficient. The type of regression (e.g. linear, multi-level) depends on the study design (within-subject vs. between-subject effect). Linear regression is used when

there is no within-subjects effect in the data, while multi-level modeling is used if there is at least one within-subject factor. Next, the analyst classifies the new requirements the experts provided into broader categories and links these to the factors/levels in the scenario. In our study, we analyzed the dependencies in the bootstrapping stage using multi-level modelling, and in the verification stage using linear regression.

The collected new expert requirements mitigations are expressed in natural language. The problem with natural language statements is that different experts could describe the same requirement using different words and phrases. As a first step, requirements are coded using short phrases (concept labels), an open coding grounded analysis approach(Hibshi et al., 2015; Saldaña, 2012). Then, the analyst categorizes the requirements using a more abstract security concept. For example, mitigations coded as password salt and stronger password, are grouped under passwords; and input sanitization and input validation are categorized under SQL injection mitigations.

After first-cycle coding and categorization, a second-cycle coding is conducted (Saldaña, 2012), where requirements are linked to the factor levels that they appear in, which would help to filter the requirements that we anticipated to appear vs. new unanticipated requirements. For example, in the network study, there are scenarios with insecure Dematerialized Zone (DMZ) configuration and a more secure split-DMZ configuration. Mitigations that suggest better network segmentation are linked to the level of the DMZ level shown in scenarios where the mitigation was elicited. If associated with the weaker DMZ, then this makes the mitigation anticipated, but if associated with the stronger DMZ, then that means there are further segmentation configurations for the network and DMZ that was not anticipated in the scenario.

In addition, each requirement is assigned one of the following codes: *refinement*, if the requirement refines the dimension by extending its functionality; a *reinforcement*, if the requirement adds auxiliary quality not directly related to the dimension; and a *replacement*, if the requirement replaces the dimension.

Upon completion of analysis, the analyst decides to either stop and be satisfied with the data collected, or continue to the next stage: *verification*. Verification is an expensive step that the analyst could pursue if the results show rich data that needs

further verification, and stop once they reach saturation. By saturation, we mean no new requirements are being collected and the analyst continues to see the same statistical results (e.g. same effect, same dependencies among the variables).

Stage 4: Verification

Based on the output of stage three, the analyst defines a selection criteria and heuristics that will guide the requirements selection process. For example, to ensure monotonically increasing quality, an analyst may only select requirements that would increase the quality of interest in the next scenarios.

In our series of security experiments, our goal is to increase security adequacy. Hence, we define the following criteria:

- For each domain, select two categories from second cycle coding with the highest number of requirements within the category.
- For each category, select the requirements with highest frequency that appear even in vignettes where the level of the requirement is strong.

In the verification stage, the requirements evaluated in the bootstrapping stage are assigned a fixed level, which is the strong security level. By fixing these levels, the effect of unanticipated requirements becomes the focus of measurement.

Then, the analyst will repeat steps from stages two and three to verify whether the new set of requirements affects the quality measurements as intended. To exit the iterative process of the MQM, the analyst establishes an end goal to be achieved. The new requirements for each of the four domains that we examined, are all shown in Table II.

TABLE II. USER STUDY SECURITY DOMAINS AND THEIR CORRESPONDING ADDED REQUIREMENT VARIABLES

Domain	Threat	Factor	Level Code	Level description
Network	Man-in-the-middle	$MFA	enabled	There is a one-time password sent to a mobile phone
			disabled	There is no further tokens or one-time passwords sent to mobile-phones
		$DBSegment	empseg	The DB Server is placed on a special admin segment separate from the employee network
			sepseg	The DB Server is placed on the same segment with the employee network
System	malware	$SWInstallation	notest	Admins are specific IT professionals who can install any new SW with no further testing
			test	New software must be tested and approved prior to installation
		$MalwareTools	enabled	Heuristic-based and behavioral-based malware-detection tools are enabled
			disabled	Heuristic-based and behavioral-based malware-detection tools are disabled
Database	Privilage escalation	$SIEM	siem	A trained IT auditor inspects logs with a specialized SIEM (Security information and event management) tool that the company installed for log analysis and management.
			nosiem	A trained IT auditor inspects logs without the assistant of costly SIEM tool
		$Notification	enabled	Admins are automatically notified when errors occur
			disabled	No notification sent to admins
Web Apps	Cross-site-request forgery	$InputValidation	client	on the client-side
			server	on the client-side, followed by input sanitization on the server-side
		$SOP	verify	In addition to the CSRF token, HTTP standard headers are examined for same origin
			noverify	The CSRF tokens are robust. No need to verify Same Origin on the server side

In our study, we select two new requirements from the reinforcement category for each security domain. The new generated scenarios will keep the bootstrapping requirements, and include new variables for the new reinforcement requirements. Since the goal is to increase security ratings, we fix the levels for the bootstrapping requirements at the strongest level. For the new requirements, we use a weak and a stronger level to test their effect in improving security ratings. Hence, each new study domain had a 2x2 factorial design (2 new variables with 2 new levels each). Table II lists all the added requirements and their levels. After deciding on the new requirements and the redesign on the new vignettes, we ran the user experiments using the same protocol from the bootstrapping stage, but with the following changes:

- Recruitment: we re-invited security analysts that we previously recruited for the bootstrapping stage and for other security-related studies by using the emails they provided *to opt-in for future studies.* We sent each participant a unique one-time code to be used to access the online survey.
- *Experiment set-up:* we set up the user experiment such that each participant sees one vignette from each domain, so the experiment has a between-subject design (no-mixed effects).

- *Statistical analysis:* since the new design is between-subject with no mixed-effect, we use linear regression for analysis

Results

Now, we will report the demographics the results of the bootstrapping and verification stages.

Demographics

In the bootstrapping stage we recruited 69 security participants. Participants had an average of 10 years of experience. The number of responses for each domain is: 39, 30, 49, and 21 for networking, operating systems, databases, and web applications, respectively (each participant was randomly assigned to two vignettes from two domains). Participants scored an average of 52% on the security knowledge test. A summary of demographics is shown in Table III below.

TABLE III. BOOTSTRAPPING STUDY: DEMOGRAPHICS

Description		Participants	
		Number	*Percentage*
Gender*	Male	59	86%
	Female	7	10%
Years of Experience* (Mean=10)	Less than 2	9	13%
	2 – 5 years	15	22 %
	6 – 10 years	15	22 %
	11 – 15 years	9	13%
	16 – 20 years	13	19%
	more than 20 years	5	7%
Job Sector*	Industry: non-research	24	35%
	Government: non-research	22	32%
	Industry: research	5	7%
	Academia	5	7%
	other	9	13%
Took academic classes in security		39	57%
Took job training in security		54	78%
Self-taught security knowledge		54	78%
Job roles	Security analyst	46	67%
	Other – IT security related	6	9%
	Other – IT related	13	19%
	Other – Non IT	4	6%
Highest Degree Completed	Bachelor's degree	31	45%
	Masters graduate degree	17	25%
	High school or equivalent	8	12%
	Some college, no degree	7	10%
	Associate degree	5	7%
	PhD degree	1	<1%
Security Knowledge Score	Scored above 60%	18	26%
	Scored between 40% and 60%	40	58%
	Scored below 40%	11	16%

* A few participants did not answer this question

The verification stage aims to evaluate to what extent the new requirements recommended in the bootstrapping stage could increase security. We sent 100 email invitations, and received 45 expert responses (45% response rate). Compared to the bootstrapping stage, respondents to the verification stage scored higher on the security knowledge test ($Mean_{Bootsrapping}$ = 52%, $Mean_{Verification}$ = 60%). In this stage, participants had an average of 9 years of experience. A demographics summary is provided in Table IV below.

TABLE IV. VERIFICATION STUDY: DEMOGRAPHICS

Description		**Participants**	
		Number	***Percentage***
Gender*	Male	43	96%
	Female	1	2%
Years of Experience* (Mean=9)	Less than 2	1	2%
	2 – 5 years	14	31 %
	6 – 10 years	16	36 %
	11 – 15 years	8	18%
	16 – 20 years	4	9%
	more than 20 years	2	4%
Job Sector*	Industry: non-research	14	31%
	Government: non-research	12	27%
	Industry: research	2	4%
	Government: research	6	13%
	Academia	3	7%
	other	7	16%
Took academic classes in security		34	76%
Took job training in security		40	89%
Self-taught security knowledge		37	82%
Job roles	Security analyst	30	67%
	Other – IT security related	4	9%
	Other – IT related	4	9%
	Other – Non IT	4	9%
Highest Degree Completed	Bachelor's degree	12	27%
	Masters graduate degree	24	53%
	High school or equivalent	2	4%
	Some college, no degree	4	9%
	Associate degree	1	2%
	PhD degree	1	2%
Security Knowledge Score	Scored above 60%	20	44%
	Scored between 40% and 60%	21	47%
	Scored below 40%	4	9%

* A few participants did not answer this question

New Requirements from the Bootstrapping Stage

Participants provided a total 550 mitigations that we classified into 55 categories and 187 sub-categories. Table V shows the top five categories for each domain based on number of occurrences (*Freq*.). The table shows how some categories appear in multiple domains (e.g. accounts/access control), while other categories were unique to a security domain (e.g. SQL injection mitigations).

TABLE V. TOP FIVE MITIGATIONS CATEGORIES

Networking		Operating Systems	
Category	***Freq.***	***Category***	***Freq.***
Passwords	29	Accounts/Access Control	59
Segmentation	20	Software Installation	21
Authentication	17	Social Media	17
Firewalls	6	Malware Detection	13
Certificates	6	White/Blacklisting	12
Databases		**Web Applications**	
Category	***Freq.***	***Category***	***Freq.***
Logs	74	Authentication	14
Accounts/Access Control	68	SQL Injection Mitigations	9
Error Handling	31	Web App Protections	9
Monitoring	10	Accounts/Access Control	4
Authentication	8	Testing	4

Comparing the Security Ratings between the Bootstrapping and Verification Stages

Our results show that the mean overall security ratings increased in the verification stage over the bootstrapping stage. This means that experts view the refined scenarios in the verification stage to have higher security adequacy than the original scenarios used in the bootstrapping stage. The results in Table VI also indicate that the average ratings are approximately 3 ±1 (STD) (adequate=3). One possible explanation could be that security experts are more conservative when rating security and cannot envision excessive security. Hibshi et al. found that security experts do prefer more conservative security ratings (Hibshi, Breaux, Riaz, et al., 2016).

The regression analysis of the verification stage also shows that the new requirements matter to the analysis, but the individual levels do not vary significantly. While in the verification stage experts report an increase in ratings over the bootstrapping stage, the increase cannot be attributed to the new requirements levels. This finding yields two key insights: security saturation, wherein it is sufficient to accept new, elicited requirements and a verification stage may not be necessary; and label bias, in which the excessive label is unreachable and thus reduces the ability to measure significant differences. For a more detailed discussions of the results, see Hibshi & Breaux (Hibshi & Breaux, 2017).

TABLE VI. COMPARISON OF EXPERTS' SECURITY RATINGS

Rating Variable Name	Bootstrapping Stage	Verification Stage
	Mean Rating	*Mean Rating*
Networking		
OverallRating	2.37	2.57
NetworkAccessRating	2.70	3.09
NetworkAuthRating	2.32	3.22
DMZRating	2.53	2.82
Operating Systems		
OverallRating	2.10	2.70
SocialMediaRating	2.60	3.13
AdminPriviligesRating	1.74	3.07
VirusScanRating	2.73	2.80
DataBases		
OverallRating	2.51	2.34
DBAccessRating	2.62	2.71
DBMonitorRating	2.56	3.00
ErrorRating	2.25	2.60
Web applicatons		
OverallRating	1.80	2.62
WebAuthRating	2.05	2.69
StoredUserDataRating	1.86	3.07

Dependency Analysis from the Bootstrapping Stage

The `$OverallRating` represents the experts' security rating of the scenario based on the composition of the requirements. We show an example of the regression equation for the web applications domain. Equation 1 is our additive regression model with a random intercept (ϵ) grouped by participant ID.

$$\$OverallRating_{webapp} = \alpha + \beta_w \$WebAuth + \beta_s \$StoredUserData + \epsilon \quad (1)$$

The additive model is a formula that defines the `$OverallRating` in terms of the intercept (α) and a series of components. Each component is multiplied by a coefficient (β) that represents the weight of that variable in the formula. The formula in Eq. 1 is simplified as it excludes the dummy (0/1) variable coding for the reader's convenience. We use the same formula for each domain, but we replace the independent variables corresponding to the factors in that domain. We follow a similar model for the individual requirements ratings. For example, Equation 2 below is the additive regression model for `$WebAuthRatings` variable.

$$\$WebAuthRating_{webapp} = \alpha + \beta_w \$WebAuth + \beta_s \$StoredUserData + \epsilon \quad (2)$$

We report the significant results of our bootstrapping stage data in Table VIII. We use the variable and level codes shown in Table I. For each security domain, we establish a baseline level for factors in that domain. The intercept (α) is the value of the dependent

variable when the independent variables are at their baseline values. The baseline levels for each domain are shown in Table VII. Table VII also shows the coefficient estimates (Coeff. Est.), which show by how much the security requirement level increased or decreased the mean rating of adequacy.

TABLE VII. SIGNIFICANT MULTILEVEL REGRESSION RESULTS FOR THE BOOSTRAPPING DATA

Dependent Variable (DV)	Independent Variable (IV) - level	Coeff. Est.	Std. Error
Networking IVs: $NetworkAccess+$NetworkAuth+$DMZ			
baseline	*offsite+ comp16 + allnosplit*		
OverallRating	*Intercept (baseline)*	*1.83****	*0.28*
	NetworkAuth - (multi8)	0.96**	0.34
Network Auth-Rating	Intercept (baseline)	2.28***	0.30
	NetworkAuth - (multi8)	0.75*	0.36
	NetworkAuth - (stand6)	-0.72*	0.36
Operating Systems IVs: $SocialMedia+$AdminPriviliges+$VirusScan			
baseline	*permit+ auth + files*		
OverallRating	*Intercept (baseline)*	*2.2****	*0.39*
	AdminPrivileges- noauth	-0.95*	0.37
SocialMediaRating	*Intercept (baseline)*	*2.06****	*0.40*
	SocialMedia- prohibit	1.13***	0.19
AdminPriviliges-Rating	*Intercept (baseline)*	*2.31****	0.43
	AdminPrivileges- noauth	-1.33***	0.41
VirusScan-Rating	*Intercept (baseline)*	*2.61****	0.35
	VirusScan - filesmemoryprocesses	0.89*	0.37
Database IVs: $DBAccess+$DBMonitor+$Error			
baseline	*extserver + available + nouser*		
OverallRating	*Intercept (baseline)*	2.89*	0.33
interaction terms	Error - user	-1.35**	0.45
	DBAccess - sqlauth * DBMonitor - month	-0.60**	0.29
	DBAccess - sqlauth * DBMonitor - needed	-0.57*	0.28
	DBMonitor - month * Error - user	1.33**	0.60
ErrorRating	*Intercept (baseline)*	2.8***	0.28
	Erroruser	-0.98***	0.27
Web Applications IVs: $WebAuth+$StoredUserData			
baseline	*basic + cescpost*		
OverallRating	*Intercept (baseline)*	2.36***	0.21
	StoredUserData - get	-0.73***	0.25
	StoredUserData - post	-1.32***	0.29
	StoredUserData - cpost	-0.70***	0.29
WebAuthRating	*Intercept (baseline)*	2.04***	0.26
	WebAuthform	0.76***	0.21

(*p≤.05 **p≤.01 ***p≤.001)

For the networking domain study, we found a significant contribution of the three network factors (`$NetworkAccess`, `$NetworkAuth`, and `$DMZ`) for predicting the `$OverllRatingNetwork` ($\chi 2$ (7) =11.3, p=0.022), over the null model (without the factors). Table VIII shows a significant effect from multifactor authentication for the network authentication requirement (coded multi8, see Table I), increasing the ratings over the intercept (1.83) by approximately one point (0.96) on the adequacy scale (almost adequate). Among all networking scenario requirements, only `$NetworkAuthRating` shows a significant effect ($\chi 2$ (4) =18.3, p=0.001) (see Table VIII).

In the database domain, we see an effect for the interaction terms of the regression model for the overall security rating ($\chi 2$ (9) =20.7, p=0.01). Reporting errors to users (Error – user) decreased the security rating by more than a point, but when the reporting errors to users are combined with a more frequent logging mechanism (DBMonitor - month) the rating increases over the baseline.

Dependency Analysis from the Verification Stage

Recall from above, the MQM uses linear regression to analyze the results of the vignette surveys responses in the verification stage. The independent variables in the regression formula are the requirements variables shown in Table II to verify the effect of the new requirements on the security ratings. We now report the regression results for each security domain.

Networking: the regression model shows that different levels of the new requirements variables `$MFA`, and `$DBSegment` do not significantly predict the `$Overall` security rating, because the regression model of `$Overall` ratings as a function of the `$MFA`, and `$DBSegment` did not show any significance over the intercept-only model (F(2,39) = 1.595, p=0.2). Hence, the `$Overall` mean, which is the intercept-only model is a better predictor of the overall security ratings for the networking study. The result is similar for the regression models constructed for the `$NetworkAccessRating`, `$NetworkAuthRating`, `$DMZRating` with: (F(2,42)=1.2, p=0.3), (F(2,42)=0.04, p=0.9) and (F(2,42)=0.5, p=0.6), respectively. The `$MFA` variable that represent multifactor authentication is shown to be a good predictor of the experts

$MFARating (F(2,42)=5.3, p<0.01). Scenarios that include multifactor authentication show an increase of 0.85±0.27 (standard error) on the *$MFARating* scale (p<0.001). Similarly, scenarios where the database is in a separate segment (*$DBSegment*) shows a significant increase (p<0.001) in the $DBSegemtnRating by 1.4±0.30 (F(2,41)=11.4, p<0.001).

Operating Systems: The regression for $Overall ratings as a function of $SWInstallation, and *$MalwareTools* show significance (F(2,41)=4.57, p=0.02) over the intercept-only model. When inspecting the coefficients, only the intercept and $MalwareTools show significant effects. Enabling heuristic-based and behavioral-based malware-detection tools show a significant increase (p=0.02) in the $Overall ratings by 0.53±0.22 and also show a significant increase (p<0.001) in the $MalwareRating by 1.68±0.16; thus, $MalwareTools is a good predictor of the $MalwareRating (F(2,42)=61.26, p<0.001). $SWInstallation is found to be good predictor of the $SWInstallationRating (F(2,42)=35.25, p<0.001). Scenarios that include testing new software prior to installation *(*$SWInstallation) show a significant increase of 1.5±0.18 of the $SWInstallationRating. We found no significant effect for the regression models constructed for the $SocialMediaRating, $AdminPriviligesRating, $VirusScanRating with: (F(2,42)=1.33, p=0.3), (F(2,42)=1.63, p=0.2) and (F(2,42)=1.45, p=0.2), respectively. We also found no significant effect for the interaction terms.

Databases: The regression model of $Overall ratings as a function of $SIEM, and $Notification show no significance (F(2,38)=1.06, p=0.35) over the intercept-only model. Except for $DBMonitorRating and $NotificationRating, no significant effects are found for the requirements ratings in the database scenarios. Database scenarios that include using a specialized SIEM (security information and event management) tool, show a significant (p=0.009) increase of 0.54 ±0.20 on the $DBMonitorRating. The $SIEM shows significance in predicting the *$DBMonitorRating* (F(2,42)=3.8, p=0.03). Similarly, $Notification is a good predictor of the $NotificationRating (F(2,42)=24.29, p<0.001). Scenarios that

include notifying admins about errors show a significant ($p<0.001$) increase of 1.48 ± 0.22 on the `$DBMonitorRating`.

Web Applications: Except for the regression model constructed for `$SOPRating`, which rates the same origin policy, no significant effects are found for the `$Overall` rating nor for all other requirements in this scenario. For the `$SOPRating`, it was not the `$SOP` variable that significantly affected this rating, but the *$InputValidation*. Scenarios that include validating the client's input on the server-side, show a significant ($p=0.007$) increase of 0.9 ± 0.32 on the `$SOPRating`. The `$InputValidation` show significance in predicting the `$SOPRating` ($F(2,39)=4.03$, $p=0.03$).

The major takeaway is that the intercept-only model is sufficient to explain the outcome dependent variable. The significance of the intercept-only model means that we can rely on using the means of the dependent variables to explain the observations in the data. For the security analyst, this means that varying levels of new factors did not show significance, but we cannot remove the factors from the model.

Threats to Validity

External validity concerns how well results generalize to the population (Shadish, Cook, & Campbell, 2002). Our target population is security experts and we recruit security professionals who attend security conferences. To assess security expertise, we measured years of experience (mean=10.0 years) and we conducted a security knowledge test that included technical questions about how to configure file permissions, network firewalls, etc.

Internal validity is the degree to which a causal relationship can be inferred between the independent and dependent variables (Shadish et al., 2002). We randomize the assignment of participants to conditions, and we randomize the presentation order of scenarios. Based on our pilot results, we limited the number of vignettes shown to four vignettes per participant to reduce fatigue. We ran the verification study seven months after the bootstrapping stage to reduce learning effects.

Construct validity is the degree to which a measurement corresponds to the construct of interest (Shadish et al., 2002). In each scenario, we present one-sentence definitions for the security level terms inadequate, adequate, and excessive, to encourage participants to interpret the label levels, similarly. The label name choice was evaluated in separate prior studies by Hibshi and Breaux (Hibshi & Breaux, 2016; Hibshi, Breaux, & Wagner, 2016).

Increasing *power* in user experiments reduces Type II errors (false negatives). We increase our power in the bootstrapping stage by using repeated measures within-subject effect, and analyzing the data with multi-level modeling, which assigns a random intercept for each subject and hence, limits the biased covariance estimates (Gelman & Hill, 2006). For a power of 80% or above, we estimate a sample size of 30 participants for the networking, operating systems, and database scenarios and 24 participants for the web applications scenario. We achieved higher sample sizes than these minimum estimates. For the verification phase, we estimate 30 participants per domain to achieve at least 80% power, and our actual sample size is 45 participants per domain.

THIS PAGE LEFT INTENTIONALLY BLANK

Discussion, Future Work and Conclusions

In this report, we explain the MQM method that provides means to empirically elicit and score security requirements from security experts. We now discuss our findings and their impact on the DoD acquisition process.

As we have pointed out in the results sections, the mean overall security had increased in the verification stage compared to the bootstrapping stage. In our study, the MQM method help us collect security ratings from domain experts and survey the experts for new mitigations that can improve the scenarios, while maintaining affordable cost. In the bootstrapping stage, we paid $1,725 in gift cards ($6.25 per scenario) to collect evaluations of 44 scenarios from 69 experts, in addition to a $600 overhead which is the cost to send the researcher to Washington D.C. We chose the gift card value based on a $50-hourly rate, which is the average rate for experts shown in our expert-salary data and in the US Bureau of Labor Statistics (U.S. Bureau of Labor Statistics., 2016). We find this cost-effective, since we can collect data in one day by flying to one conference venue and with little effort to convince experts to participate. The data analysis took one month, and the surveys for the verification stage were completed in two weeks ($1,425 in gift cards). For an organization to hire security experts to evaluate scenarios or to perform risk-based security analysis, the cost will be more than paying the average hourly rate, due to the added overhead of experts' recruitment and accommodations. This method can be applied to collect ratings from experts on various scenarios.

The MQM employs vignette surveys to link requirements as factors to a system quality, and to elicit expert judgments about quality levels achieved by those requirements. This is different from prior work in scenario-based requirements elicitation that employs interviews (Potts et al., 1994; Sutcliffe, 1998; Van Lamsweerde, 2000). Although interviews provide detailed scenario descriptions, our approach allows analysts to attribute a quality level to specific requirements and their interactions. The MQM does not measure coverage, but it offers increased coverage of scenarios as it allows the manipulation of descriptions, and the measurement effects of certain requirements on the outcome as well as the dependencies between the requirements. In addition, the use of

surveys makes it more convenient to recruit more stakeholders, which increases the number of viewpoints of the scenario, and multiple viewpoints improve inter-personal uncertainty; which means, one expert might point out something that other experts missed, while other experts find something different. This uncertainty among experts, which impacts security assessments (Hibshi et al., 2015; Hibshi, Breaux, Riaz, et al., 2016; Hibshi, Breaux, & Wagner, 2016), is due to differences in background or human memory limitations (Hibshi, Breaux, & Wagner, 2016).

Going forward, our future research involves automating the process and building tools that leverage online cloud platforms to collect these expert ratings. Using our tool, an analyst would be able to build their own scenario and then send out invitations for experts to rate the overall security, the individual security requirements, and provide further requirements that can enhance the ratings. We envision that such a tool would have a great impact the DoD and other organizations in the public and private sectors, because it will help systemize the evaluation of security components using real-experts input. Another application includes designing digital security advisors using rules derived from the results of the vignettes.

References

Alexander, C. S., & Becker, H. J. (1978). The use of vignettes in survey research. *Public Opinion Quarterly*, *42*(1), 93–104.

Auspurg, K., & Hinz, T. (2014). *Factorial Survey Experiments* (Vol. 175). SAGE Publications. Retrieved from https://books.google.com/books?hl=en&lr=&id=1jieBQAAQBAJ&oi=fnd&pg=PP1&dq=factorial+survey+experiments&ots=Xyrl2KDAP8&sig=lGNaloueXGgXGSRRAI6LsouSqp0

Ben-Asher, N., & Gonzalez, C. (2015). Effects of cyber security knowledge on attack detection. *Computers in Human Behavior*, *48*, 51–61. https://doi.org/10.1016/j.chb.2015.01.039

Chen, T. M., & Abu-Nimeh, S. (2011). Lessons from stuxnet. *Computer*, *44*(4), 91–93.

Chi, M. T. (2006). Two approaches to the study of experts' characteristics. *The Cambridge Handbook of Expertise and Expert Performance*, 21–30.

Ericsson, K. A., & Lehmann, A. C. (1996). Expert and exceptional performance: Evidence of maximal adaptation to task constraints. *Annual Review of Psychology*, *47*(1), 273–305.

Garfinkel, S. L. (2012). The Cybersecurity Risk. *Commun. ACM*, *55*(6), 29–32. https://doi.org/10.1145/2184319.2184330

Garlan, D. (2010). Software engineering in an uncertain world. In *Proceedings of the FSE/SDP workshop on Future of software engineering research* (pp. 125–128). ACM. Retrieved from http://dl.acm.org/citation.cfm?id=1882389

Gelman, A., & Hill, J. (2006). *Data analysis using regression and multilevel/hierarchical models*. Cambridge University Press. Retrieved from https://books.google.com/books?hl=en&lr=&id=c9xLKzZWoZ4C&oi=fnd&pg=PR17&dq=data+analysis+using+regression&ots=baO8OZKpmf&sig=OViqOljyuZN_6OVZ6ybPZSJpIjA

Gomez, M. A., & Villar, E. B. (2018). Fear, Uncertainty, and Dread: Cognitive Heuristics and Cyber Threats. *Politics and Governance*, *6*(2), 61–72. https://doi.org/10.17645/pag.v6i2.1279

Goodall, J. R., Lutters, W. G., & Komlodi, A. (2004). I know my network: collaboration and expertise in intrusion detection. In *Proceedings of the 2004 ACM conference on Computer supported cooperative work* (pp. 342–345). ACM. Retrieved from http://dl.acm.org/citation.cfm?id=1031663

Goodall, J. R., Lutters, W. G., & Komlodi, A. (2009). Developing expertise for network intrusion detection. *Information Technology & People*, *22*(2), 92–108.

Graham, I. (1996). Task scripts, use cases and scenarios in object oriented analysis. *Object Oriented Systems*, *3*(3), 123–142.

Hardee, J. B., West, R., & Mayhorn, C. B. (2006). To download or not to download: An examination of computer security decision making. *Interactions*, *13*(3), 32–37.

Hibshi, H., Breaux, T., & Broomell, S. B. (2015). Assessment of Risk Perception in Security Requirements Composition. *2015 IEEE 23rd International Requirements Engineering Conference (RE)*, 146–155.

Hibshi, H., & Breaux, T. D. (2016). *Evaluation of Linguistic Labels Used in Applications* (Technical Report). Carnegie Mellon University.

Hibshi, H., & Breaux, T. D. (2017). Reinforcing Security Requirements with Multifactor Quality Measurement. In *2017 IEEE 25th International Requirements Engineering Conference (RE)* (pp. 144–153). Lisbon, Portugal: IEEE.

Hibshi, H., Breaux, T. D., Riaz, M., & Williams, L. (2016). A Grounded Analysis of Experts' Decision-Making during Security Assessments. *Journal of Cybersecurity*. Retrieved from http://cybersecurity.oxfordjournals.org/content/early/2016/10/04/cybsec.tyw010.abstract

Hibshi, H., Breaux, T. D., & Wagner, C. (2016). Improving security requirements adequacy: an interval type 2 fuzzy logic security assessment system. In *2016 IEEE Symposium Series on Computational Intelligence (SSCI)* (pp. 1–8). IEEE. Retrieved from http://ieeexplore.ieee.org/abstract/document/7849906/

Jacobson, I. (1992). *Object Oriented Software Engineering: A Use Case Driven Approach* (1 edition). New York : Wokingham, Eng. ; Reading, Mass: Addison-Wesley Professional.

Johnston, A. C., Warkentin, M., McBride, M., & Carter, L. (2016). Dispositional and situational factors: Influences on information security policy violations. *European Journal of Information Systems*, *25*(3), 231–251.

Letier, E., Kramer, J., Magee, J., & Uchitel, S. (2005). Monitoring and control in scenario-based requirements analysis. In *Proceedings. 27th International Conference on Software Engineering, 2005. ICSE 2005.* (pp. 382–391). https://doi.org/10.1109/ICSE.2005.1553581

Lincoln, Y. S., & Guba, E. G. (1985). *Naturalistic inquiry* (Vol. 75). Sage. Retrieved from https://books.google.com/books?hl=en&lr=&id=2oA9aWlNeooC&oi=fnd&pg=PA7&dq=+Naturalistic+inquiry&ots=0spzWfPeAt&sig=dhDiu_facEnd9ijVfrnUnr8lIIM

Lipshitz, R., & Strauss, O. (1997). Coping with Uncertainty: A Naturalistic Decision-Making Analysis. *Organizational Behavior and Human Decision Processes*, *69*(2), 149–163. https://doi.org/10.1006/obhd.1997.2679

Liu, L., Yu, E., & Mylopoulos, J. (2002). Analyzing security requirements as relationships among strategic actors. In *Submitted to the Symposium on Requirements Engineering for Information Security (SREIS'02), Raleigh, North Carolina*. Retrieved from ftp://www.cs.toronto.edu/cs/ftp/pub/eric/eric/SREIS02-Sec.pdf

McBride, M., Carter, L., & Warkentin, M. (2012). Exploring the role of individual employee characteristics and personality on employee compliance with cybersecurity policies. *RTI International-Institute for Homeland Security Solutions*.

McDermott, J., & Fox, C. (1999). Using abuse case models for security requirements analysis. In *Computer Security Applications Conference, 1999. (ACSAC '99) Proceedings. 15th Annual* (pp. 55–64). https://doi.org/10.1109/CSAC.1999.816013

McKelvie, A., Haynie, J. M., & Gustavsson, V. (2011). Unpacking the uncertainty construct: Implications for entrepreneurial action. *Journal of Business Venturing*, *26*(3), 273–292.

Mendel, J. M. (2001). *Uncertain rule-based fuzzy logic systems : introduction and new directions*. Prentice Hall PTR,.

Michels, S., Velikova, M., Hommersom, A., & Lucas, P. J. F. (2013). A Decision Support Model for Uncertainty Reasoning in Safety and Security Tasks. In *2013 IEEE International Conference on Systems, Man, and Cybernetics* (pp. 663–668). https://doi.org/10.1109/SMC.2013.118

Mouratidis, H., & Giorgini, P. (2007). Secure tropos: A security-oriented extension of the tropos methodology. *International Journal of Software Engineering and Knowledge Engineering*, *17*(02), 285–309.

National Institute Standards and Technology (NIST). (2018). *Supporting the Growth and Sustainment of the Nation's Cybersecurity Workforce*. National Institute Standards and Technology. Retrieved from https://www.nist.gov/itl/applied-cybersecurity/nice/resources/executive-order-13800/supporting-growth-and-sustainment

NIST/ITL Special Publication (800). (2015, January 2). Retrieved January 2, 2015, from http://www.itl.nist.gov/lab/specpubs/sp800.htm

Nuseibeh, B., Kramer, J., & Finkelstein, A. (1994). A framework for expressing the relationships between multiple views in requirements specification. *IEEE Transactions on Software Engineering*, *20*(10), 760–773. https://doi.org/10.1109/32.328995

Potts, C. (1997). Requirements models in context. In *, Proceedings of the Third IEEE International Symposium on Requirements Engineering, 1997* (pp. 102–104). https://doi.org/10.1109/ISRE.1997.566847

Potts, C., & Newstetter, W. C. (1997). Naturalistic inquiry and requirements engineering: reconciling their theoretical foundations. In *, Proceedings of the Third IEEE International Symposium on Requirements Engineering, 1997* (pp. 118–127). https://doi.org/10.1109/ISRE.1997.566849

Potts, Colin, Takahashi, K., & Anton, A. I. (1994). Inquiry-based requirements analysis. *IEEE Software*, *11*(2), 21–32.

Rossi, P. H., & Nock, S. L. (1982). *Measuring Social Judgments: The Factorial Survey Approach*. SAGE Publications.

Saldaña, J. (2012). *The coding manual for qualitative researchers*. Sage.

SANS Institute: About. (2017, February 4). Retrieved February 4, 2017, from https://www.sans.org/about/

Schmidt, F. L., & Hunter, J. E. (1993). Tacit knowledge, practical intelligence, general mental ability, and job knowledge. Retrieved from http://psycnet.apa.org/psycinfo/1993-32165-001

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference.* Houghton, Mifflin and Company.

Sutcliffe, A. (1998). Scenario-based requirements analysis. *Requirements Engineering*, *3*(1), 48–65.

Swenson, G. (2009, June 11). NIST, DOD, Intelligence Agencies Join Forces to Secure U.S. Cyber Infrastructure [Text]. Retrieved March 30, 2017, from https://www.nist.gov/news-events/news/2009/06/nist-dod-intelligence-agencies-join-forces-secure-us-cyber-infrastructure

Trinkle, B. S., Crossler, R. E., & Warkentin, M. (2014). I'm game, are you? Reducing real-world security threats by managing employee activity in online social networks. *Journal of Information Systems*, *28*(2), 307–327.

U.S. Bureau of Labor Statistics. (2016, March 8). Information Security Analysts : Occupational Outlook Handbook: U.S. Bureau of Labor Statistics. Retrieved March 8, 2016, from http://www.bls.gov/ooh/computer-and-information-technology/information-security-analysts.htm

Van Lamsweerde, A. (2000). Requirements engineering in the year 00: a research perspective. In *Proceedings of the 22nd international conference on Software engineering* (pp. 5–19). ACM. Retrieved from http://dl.acm.org/citation.cfm?id=337184

Wallander, L. (2009). 25 years of factorial surveys in sociology: A review. *Social Science Research*, *38*(3), 505–520. https://doi.org/10.1016/j.ssresearch.2009.03.004

Acquisition Research Program
Graduate School of Business & Public Policy
Naval Postgraduate School
555 Dyer Road, Ingersoll Hall
Monterey, CA 93943

www.acquisitionresearch.net