JOURNAL OF CYBERSECURITY

Journal of Cybersecurity, 2016, 1–17
doi: 10.1093/cybsec/tyw010
Research paper

Research paper

A grounded analysis of experts' decision-making during security assessments

Hanan Hibshi[1,*], Travis D. Breaux[1], Maria Riaz[2] and Laurie Williams[2]

[1]Institute for Software Research, Carnegie Mellon University, Pittsburgh, PA 15213, USA

[2]Department of Computer Science, North Carolina State University, Raleigh, NC 27695, USA

*Corresponding author: E-mail: hhibshi@cs.cmu.edu

Received 13 February 2016; revised 8 July 2016; accepted 15 August 2016

Abstract

Security analysis requires specialized knowledge to align threats and vulnerabilities in information technology. To identify mitigations, analysts need to understand how threats, vulnerabilities, and mitigations are composed together to yield security requirements. Despite abundant guidance in the form of checklists and controls about how to secure systems, evidence suggests that security experts do not apply these checklists. Instead, they rely on their prior knowledge and experience to identify security vulnerabilities. To better understand the different effects of checklists, design analysis, and expertise, we conducted a series of interviews to capture and encode the decision-making process of security experts and novices during three security analysis exercises. Participants were asked to analyze three kinds of artifacts: source code, data flow diagrams, and network diagrams, for vulnerabilities, and then to apply a requirements checklist to demonstrate their ability to mitigate vulnerabilities. We framed our study using Situation Awareness, which is a theory about human perception that was used to elicit interviewee responses. The responses were then analyzed using coding theory and grounded analysis. Our results include decision-making patterns that characterize how analysts perceive, comprehend, and project future threats against a system, and how these patterns relate to selecting security mitigations. Based on this analysis, we discovered new theory to measure how security experts and novices apply attack models and how structured and unstructured analysis enables increasing security requirements coverage. We highlight the role of expertise level and requirements composition in affecting security decision-making and we discuss how our method produced new hypotheses about security analysis and decision-making.

Key words: security; requirements; patterns; analysis; decision-making; ysituation awareness

Introduction

Each year, attackers exploit well-known vulnerabilities that have obvious, well-documented solutions. Hewlett-Packard's top cyber security risks report in 2011 presents many popular attacks against web applications, such as SQL injection attacks [1]. In addition, the Open Web Application Security Project (OWASP) top 10 web application security vulnerabilities [2] and the SANS top 20 critical security controls [3] aim to reduce the most common vulnerabilities. Finally, high-profile standards bodies publish security control catalogues, including the ISO/IEC 27000 Series standards and the US National Institute of Standards and Technology (NIST) Special Publication 800 Series that contain best practice security requirements [4]. Despite these broadly disseminated, diverse, and in-depth sources of security knowledge, information systems continue to be susceptible to known vulnerabilities. Many systems continue to operate under poor security practices, such as unencrypted wireless networks, using the same administrative password across multiple systems, which are unexpired and outdated [5].

We can refer to security as a "wicked problem." Wicked problems could be defined as those difficult to solve because of unclear,

ambiguous, or conflicting requirements [6,7] making possible solutions difficult to be enumerated [8]. For example, security analysts may respond differently to the same security problem by resolving discrepancies represented in the problem differently. With such wicked problems, researchers suggest that the design of solutions should be aimed at reducing ambiguity by coming to a collective understanding of the problem representation [7,8]. Because of this "wicked" nature, researchers suggest that security requirements could only be satisficed as opposed to satisfied; because we cannot guarantee eliminating risk but we can reduce it to an acceptable limit, which makes the term security assurance to be more acceptable as we cannot provide absolute security [9].

The lack of information system security is unlikely due to an absence of security requirements or analysis methods, which are abundant. Research in requirements engineering has sought to address security, including abuse and misuse cases [10,11], anti-goals [12], and trust assumptions that are used to construct assurance arguments [13,14]. Combined with the wealth of available security knowledge, we hypothesize insecure information systems persist because security analysts experience two challenges: (i) they experience difficulty in perceiving relevant risks in the context of their information system designs to select appropriate security requirements especially when choices of mitigations is affected by dependencies and priorities that exist among the requirements, which we call requirements composition; and (ii) they experience difficulty in deciding which requirements are appropriate to minimize risk in light of these dependencies. We propose that requirements analysis methods evaluation should address these difficulties directly; however, we have little insight into the technical challenges of designing methods to achieve this goal. Therefore, in this article, we examine different security analysts' responses to the same artifacts with and without checklists, a prominent requirements analysis, and documentation method [4,15]. The contributions of this article are as follows:

- A novel coding method to apply Situation Awareness (SA) to interview data, which we apply in security analysis to understand how security experts decide on security requirements.
- New hypothesis based on SA decision-making patterns to measure how attack models enhance security analysis and how novices and experts differ in the application of these models under uncertainty.
- New evidence based on SA decision-making patterns that explain the issues by using checklists.
- New hypotheses about security requirements composition that impact security analysis and decision-making.

The remainder of this article is organized as follows: we present background on situation awareness in section "Situation awareness and security risk"; our research method in section "Research approach"; results of evaluating our approach in section "Evaluation of approach"; the decision-making patterns in section "Decision-making patterns"; discuss our observations of participant expertise in section "Participants' expertise and the attacker model"; threats to validity in section "Observations across the three artifact categories"; followed by our discussion in section "Threats to validity." Finally, we conclude in section "Discussion."

Situation Awareness and security risk

"Situation Awareness" (SA) is framework introduced by Mica R. Endsley in 1988 [16] that distinguished between a user's "*perception* of the elements in the environment within a volume of time and space, the *comprehension* of their meaning, and the *projection* of their status in the near future" during their engagement with a system. Perception, comprehension, and projection are called the levels of SA, and a person ascends through these levels in order to reach a decision. To illustrate, consider SQL injection, in which an attacker inserts an SQL statement fragment into an input variable (often via a web form) to gain unauthorized database access. When an expert conducts a source code (SC) vulnerability assessment, they look for cues in the code to place input sanitization, which is a mitigating security requirement. Upon finding such cues (perception), analysts proceed to reason about whether the requirement has or has not been implemented (comprehension). Once understood, they can informally predict the likelihood of an SQL injection attack and the consequences on the system (projection) based on their experience and understanding of the threat and attack vector.

We believe SA can be used to explain how analysts perform risk assessments. The NIST Special Publication 800-30 [4] defines risk as the product of the likelihood that a system's vulnerability can be exploited and the impact that this exploit will have on the system. The ability to predict likelihood and impact depend on the analyst's ability to project prospective events based on what they have perceived and comprehended about the system's specification and its state of vulnerability. If the expert succeeds in all three SA levels, then they have "good" SA and they should be able to make more accurate decisions about security risks. Failure in any level results in "poor" SA that leads to inaccurate decisions or no decisions at all. In section "Research approach," we describe our method to detect the SA levels in security expert interviews.

Endsley and other researchers [16–18] go beyond the SA definition to establish a holistic framework that scientists in other fields could benefit from and apply. This framework entails details and relationships to other concepts such as: expertise effect, goals, mental models, automation, uncertainty, requirements analysis, etc. A schema is a known key term in cognitive psychology defined as the mental framework in the human's cognition of prepossessed ideas that represent some aspects of the world. Schemata are a group of schemas organized in cognition that improve humans' ability to retrieve knowledge or acquire new knowledge [19–21]. For example, when we solve new problems using a computer programming language, schema theory suggests that our cognition matches the new problem structure with existing schemata for solving past problems and this process is what cognitive psychologists call: schema abstraction [22]. Rao *et al.* found that the number and variety of training examples in programming language experiments had minimal effect on schema abstraction [23]. Thus, we may conclude that schema abstraction is an expert ability that is acquired over multiple, repetitive examples across different contexts. Endsley explains how expertise can help a person build and enhance mental schemata which in turn, facilitates the person's ability to interpret their perceptions and make necessary projections that lead to better decisions [17].

The SA framework is flexible and could be customized according to the needs of a system. Examples of fields in which SA has been applied include military operations [24], command and control [25], cyber security, [26] and many others [17,27]. Researchers have modeled SA in intelligent and adaptive systems [24,25,27]. Feng *et al.* proposed a context-aware decision support system that models situation awareness in a command–control system [25]. Their focus was to have entity agents based on a "rule-based inference engines" that provide decision support for users. They applied Endsley's concepts and focused on "Shared Situation Awareness" along with a computational model that they applied to a case study of a command and control application. Chen *et al.* extended a cyber

intrusion detection system using a formalization of SA concepts; the logic formalization is derived from experts' experiences [26]. Jakobson proposed a framework of situation-aware multiagent systems that could be cyberattack tolerant [28]. To our knowledge, SA has not been widely adopted in security requirements engineering.

Research approach

We chose an exploratory, qualitative research method that aims to understand the symbolic and cognitive processes of specific security analysts, as opposed to testing hypotheses against specific variables [29]. The purpose of our approach is to develop a theory of security analysis from a rich dataset that we can later test in a controlled experiment. We are interested in how security analysts make decisions, and whether their decisions lead to optimal solutions. Consequently, this theory is grounded in the domain and findings from this study are only generalizable for this dataset [29]. Our method consists of three main phases:

- The "preparation phase," in which we developed the research protocol, including tailoring SA to security analysis, selecting the system artifacts to use in the analysis, and recruiting the security analysts to be interviewed;
- The "interview phase," wherein we elicited responses from the selected analysts; and
- The "qualitative data analysis phase," in which we coded the interview transcripts and systematically drew inferences from the data.

We employed coding theory [30] to link SA concepts to the dataset and validate whether our observations are consistent and complete with respect to that dataset. In Glaser's view of grounded theory, investigators approach the dataset without any preconceived concepts to explain the data, also called open coding [38]. In our approach, we adopted Strauss' view [29] that begins with an initial coding frame, called hypotheses coding [30], that we based on Endsley's SA levels. We employed a two-cycle coding method: in the first cycle, we applied the "hypothesis coding" method to our dataset using the predefined code list derived from Endsley's SA levels. This method tests the validity of the initial code list. In the second cycle, we applied theoretical coding to discover decision-making patterns from the dataset [30]. We now discuss the three phases.

The preparation phase

The SA framework can be tailored to a field of interest by mapping SA levels to statements made by domain analysts. We tailored the framework by verbally probing the analyst during the interview process as they were asked to evaluate security risk of information system artifacts. We expected the dataset to show how analysts build situation awareness. We also expected it to help us further discover how perceptions of security risk evolve as the analysts' awareness of both potential vulnerability and available mitigations increases. The inability to perceive risk may be due to limitations in analysts' knowledge or ambiguities in the artifacts. We define the SA levels as follows:

- Level 1: Perception: the participant acknowledges perceiving security cues in the given artifact. Examples include: "there is a picture of a firewall here" or "there are SQL commands in the code snippet." Each observation excludes any deeper interpretation into the meaning of the perception.
- Level 2: Comprehension: the participant explains the meaning of cues that they perceived in Level 1. They provide synthesis of perceived cues, analysis of their interpretations, and comparisons to past experiences or situations. Examples of comprehension include: "the firewall will help control inbound and outbound traffic…" and "the SQL commands are used to access the database which might contain private information, so we need to check the input to those commands, but this is not done in the code…"
- Level 3: Projection: the participant has comprehended sufficient information in Level 2, so they can project future events or consequences. In security, projections include potential, foreseeable attacks, or failures that result from poor security. Examples include: "this port allows all public traffic, which makes the network prone to attacks…", or "unchecked input opens the door to SQL injection…"

Finally, after Level 3, we expect participants can make security-related decisions. Decisions include steps to modify the system to mitigate, reduce, or remove vulnerabilities. Continuing with the SQL injection example, one decision could be: "this port should be closed" or "a function should be added here that checks the input before passing it to the SQL statement". Closing the port prevents an attacker from exploiting the open port in an attack, whereas checking the input can remove malicious SQL in an SQL-injection attack.

Security artifacts

We presented each participant with three categories of security-related artifacts: SC, data flow diagrams (DFD), and network diagrams. We chose these artifacts to cover from low-level SCs to high-level architecture, noting that security requirements should be mapped to each artifact in different ways and analysts require different skills to do this mapping. Based on our own experience and knowledge of security expertise, we considered the effect of specialization in areas such as secure programming, network security, etc. in selecting these artifacts. Hence, the selection aims to satisfy two goals: (i) to account for diverse background and experience; and (ii) to assess whether different artifacts show differences among SA levels. We discuss our analysis results to address these two goals in sections "Participants' Expertise and the Attacker Model" and "Observations across the three artifact categories." We now describe the artifacts used in this study.

Source Code (SC). We present participants with JavaScript code snippets, corresponding SQL statements, and a picture of a web user interface related to the snippet. The SC contains two vulnerabilities: an SQL injection attack and unencrypted username and password. JavaScript is a subset of a general purpose programming language, i.e. no templates, pointers, or memory management. Thus, we expect analysts with general programming language proficiency and knowledge of SQL injection to be able to spot these vulnerabilities in the SC. We also list a high-level security goal to prompt participants and we ask participants if the goal has been satisfied.

Data Flow Diagram (DFD). We present participants with a DFD for installing an application on a mobile platform. As shown in Fig. 1, the diagram contains high-level information about the data flow between the user, app developer, and the market. The participants are asked about possible security requirements to ensure secure information flow, and whether they can evaluate those requirements based on this diagram.

Network Diagrams (ND). We present to participants two network diagrams: ND1 shows an insecure network, and ND2 shows a network with security measures that address weaknesses in ND1. After participants are provided time to study ND1, we present ND2 and ask participants to evaluate whether ND2 is an

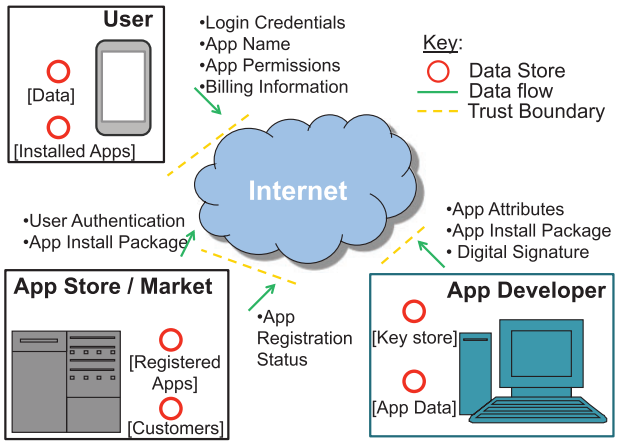

Figure 1. The data flow diagram artifact.

improvement over ND1. After collecting data on participants' evaluation of ND2, we present 15 security requirements to participants, which we explain are part of a security improvement process, and we ask participants to assess whether the network in ND2 satisfies the 15 requirements (shown in Appendix).

All of the selected artifacts are typical examples comparable to what is generally taught in college-level security courses. For example, the network diagrams were originally used in the Applied Information Assurance Class taught by Christopher May at Carnegie Mellon University [31].

Selecting experts for the study

In this study, we aim to observe how security expertise affects requirements analysis. However, security analysts are not all equal in expertise; some analysts have more experience than others in particular areas, and training in academia is different than hands-on practice. To cover a broad range of expertise, we invited industrial practitioners and PhD students at different stages of matriculation, all working in security. We present demographic data to characterize their experience levels in section "Participants' background and expertise."

The interview phase

We designed the interviews so as to study how analysts reach a security-related decision, and not study the correctness of the decision or degree of security improvement. We chose this design to reduce participants' anxiety about being personally evaluated. During our interviews, we only ask the following kinds of questions:

- What cues did the participant look at? (Perception)
- How were the cues interpreted? (Comprehension)
- Why did they interpret a cue that way? (Comprehension)
- What are the future consequences of each interpretation? (Projection)
- Based on those projected consequences, what is the best practice? (Decision)

Our approach differs from how SA is traditionally studied in human operator environments (e.g. airplane cockpits and nuclear power plants) that use the Situational Awareness Global Assessment Technique (SAGAT) [17], in that our participants are not immersed in a simulation *per se*. Rather, we present artifacts (SC, DFD, ND) to participants with prompts to evaluate artifacts for vulnerabilities asking them to act as the security analyst in this setting. We observe their ability to conduct requirements analysis, their proposed modifications or decisions, and their evaluation of security requirements' satisfaction.

In addition, we ask participants to share information about their decision-making, such as unstated assumptions and what artifact cues led participants to reach a decision. We were careful not to guide participants in a particular direction by keeping our questions general. In addition, we avoided questions such as: what do you "perceive," "comprehend," or "project?" For example, if a participant identified an attack scenario, we would follow with "why would you think such an attack could occur," or "could you describe how it could happen?" Based on our approach to limit our influence on their responses, we found participants returning to the artifact to identify cues and to explain their interpretation.

Given our interest in distinguishing novices from expert analysts, we asked participants to provide a brief description of their relevant background. Questions to elicit background information were asked twice: first, at the interview start, we ask participants about their security background, their education, industry experience, and security topics of interest. Lastly, at the end, we ask the participant about the analysis process they used during the interview and how it relates to their background. Finally, we recorded the interviews for transcription and analysis.

The qualitative data analysis phase

Grounded analysis is used to discover new theory and to apply existing theory in a new context [29]. We apply grounded analysis in three steps: (i) we transcribe the interviews; (ii) beginning with our initial coding frame (see Table 1), we code the transcripts by identifying phrases that match our codes, while discovering new codes to further explain phrases that do not match our preconceived view of the data; and (3), we review previously coded datasets to ensure the newly discovered codes were consistently applied across all transcripts. After the pilot, we observed uncertainty among participants so we added codes to capture the uncertainty. Table 1 shows the complete coding frame: the first eight codes (P, C, J, D, including the variants that account for uncertainty U*) constitute the initial coding frame and were inspired by Endsley's terminology for the Situation Awareness [17]; the remaining four codes were discovered during our analysis to account for the interview mechanics. We employed two coders (the first and third authors) who first met to discuss the coding process and coding frame, before separately coding the transcripts, and finally meeting to resolve disagreements. The process to resolve disagreements led to improvements in the form of heuristics that explain when to choose one code over the other in otherwise ambiguous situations. To efficiently identify disagreements, we used a fuzzy string-matching algorithm [32] to align the separately coded transcripts. Finally, each coder recorded their start and stop times.

To ensure all statements are coded, we applied the null code {NA} to any statements that did not satisfy the coding criteria, such as when participants request a scrap piece of paper to draw a figure, or when they ask how much time is remaining for the interview, and so on. We code statements, such as: "I took a course in security…" or "I saw on the news a security breach related to this artifact" as background {BG}, which includes their personal experience and knowledge. If the participant compares and contrasts comprehended information from the artifact to their experience or knowledge, then that information is coded as comprehension {C}. To improve construct validity, the two raters resolved borderline cases by discussing and refining the code definitions and heuristics. The following

Table 1. Situation awareness annotation codes

Code name and acronym	Definition and coding criteria used to determine applicability of the code
Perception {P}	Participant is acknowledging that they can see certain cue(s)
Comprehension {C}	Participant are explaining the meaning of cue(s) and conducting some analysis on the data perceived
Projection {J}	Participant is predicting possible future consequence(s) or risk(s) involved
Decision {D}	Participant is stating their decision
Uncertain Perception {UP}	Uncertainty at perception level: participant is missing certain data that would help they need to analyze the artifact
Uncertain Comprehension {UC}	Uncertainty at comprehension level: participant is not missing data but they cannot interpret their meaning confidently
Uncertain Projection {UJ}	Uncertainty at projection level: participant cannot predict possible future consequences confidently
Uncertain Decision {UD}	Uncertainty in decision: participant is not confident about the decision that should be made
Assumption {A}	Participant is stating assumption(s)
Ask Question {Q}	Participant is asking the interviewer questions
Probe {Pro}	Interviewer is triggering the participant's thinking with questions or guidance information
Background {BG}	Participant is providing information regarding their personal background
Null code {NA}	Statement is not applicable to code criteria above

heuristics were used to classify statements and draw clearer boundaries between coded data:

- **Perception**: The participant verbally identifies a cue in the data (e.g. line number in code, an entity on the network diagram, a specific requirement in the text). Participants are only reporting what they see, and are not commenting or analyzing the cue.
- **Comprehension**: The participant analyzes, makes inferences, or makes comparisons about what they see. This may include the name of the cue (e.g. firewall), but the statement at least includes an interpretation in addition to reporting the perception of the cue.
- **Projection**: The participant forecasts future attacks, possible threats, or any events that could occur based on the context found in the artifact.
- **Decision**: The participant makes a decision with regard to the context. This includes deciding whether the system is secure or not, or if a certain requirement is satisfiable. Introducing new mitigations of security threats are also considered decisions.
- **Uncertainty (at any SA level)**: To determine if the participant is uncertain, first examine the verbal cues that indicate uncertainty, including, but not limited to: "I guess", "I am not sure", and "this is not clear to me." For example, the participant may indicate that they do not know what an icon represents. Alternatively, if the participant acknowledges that they see a cue, but that they cannot understand its role in the artifact, then this is an uncertain comprehension. Uncertainty can also be a cause by missing cues: a participant might indicate that they do not see a certain cue that they need for the analysis, which we interpret as uncertain perception.
- **Assumption**: The participant here needs to overtly express that they are making an assumption. Examples of such statements include: "I am going to guess that this means", "I assume", "Based on my experience this means, but it's not necessarily what the artifact tells me," and so on. To clarify how to distinguish assumptions from comprehensions, a comprehension is when the participant is explaining a certain cue's meaning based on the information given in the artifact. Assumptions, however, provide further explanation based on the participant's experience with similar systems to compensate for missing cues or missing information in the artifact.

After the first cycle coding, we conducted a second cycle or axial coding [30] to identify decision-making patterns. We defined cutoffs between coded sequences by sequentially numbering each statement and then assigning group numbers to statements that address the same expanding idea. The groups serve to delineate transitions between units of analysis. We programmatically extracted SA-level sequences (e.g. P-C for perception followed by comprehension) that we later associated with separate, named patterns, and we searched the dataset without the cutoffs to assess pattern validity (i.e. detect false positives: does the SA-level sequence always correspond to an actual coherent pattern that we assigned?). We recorded false positives in which the sequence appeared in the data, but did not conform to the pattern. We used the false positives identification to compute pattern accuracy or ratio of true positives over the sum of true and false positives.

The next step in our grounded analysis includes labeling interviewee statements with entity identifiers from the specifications, such as variables and functions in the SC or servers and firewalls in the network diagram. Once labeled, we were able to sort our analysis results by entity to see how different participants react to and analyze the same entity and to link the decision patterns to corresponding entities. We report the results of the entity analysis under section "Participants' Expertise and the Attacker Model" with respect to the role of attack models. We also report the results in section "Observations across the three artifact categories" to reflect on participants' performance among the different artifact types.

Pilot study

We piloted the study on two experts: participant P1 is an expert with extensive hands-on and academic expertise in networks and systems security; and participant P2 is a novice who has only academic security experience. The purpose of the pilot study is to test our interview protocol and apply any needed modifications to the questions or protocol before conducting additional interviews.

Reliance on assumptions and searching for more information are both uncertainty resolution techniques that are explained in Endsley's work [17]. However, it is interesting to see in our pilot results that experts and novices apply the techniques differently. Both participants P1 and P2 analyzed the network diagram artifact, but P2 was unable to think deeper about certain details and reported a higher number of uncertainties. One insight that we observed in the pilot study was the ability of the more experienced participant P1 to make assumptions when faced with uncertainty. When the novice participant was faced with uncertainty, their solution was to ask the interviewer clarification questions. The following excerpt below is

an example of an assumption that participant P1 made when they analyzed the requirement R9 that states implementing time synchronization for logging and auditing capabilities. Note that each statement will have an opening and closing code tags (see Table 1 for codes):

> {UP}I don't see an NTP server on this network{/UP} {C}but I know that Windows Domain Controller can act as NTP{/C}, {A}so I am going to assume that when they install it they'll probably leave that box checked because it's a default option{/A}.
>
> {D}I think that is probably happening here{/D}

When P2 was faced with uncertainty, however, they turned to the interviewer and asked:

> {Q} What kind of software does this thing has? {/Q}

An observation during our pilot study is that, although we asked participants to verify security requirements to check consistency between the requirements and the network diagram present in the artifact, they actually performed requirements validation, where they assess if the requirements actually meets the stakeholders' system security goals). An explanation may be that security experts rely on background knowledge and apply known security requirements. In addition, we found experts often add missing requirements, explain how to apply a requirement, evaluate whether a requirement was feasible, list some needed specifications, and prioritize requirements. For example, consider the following excerpt as participant P1 is evaluating R2 in the context of diagram ND2 and pointing out that this requirement is less critical than requirement R1 that they had evaluated earlier:

> {C}but I don't think it's as critical as say the DMZ one, but I think its sort of whatever is the next tier of criticality{/C}.

Based on our pilot study experience and the participant feedback, we revised our study protocol. A major change was the order of the presentation of network diagrams ND1 before ND2, and asking participants to draw on ND1 to improve this diagram. After this modified step, we show participants the secure diagram ND2 and ask them to compare this diagram to their own solution to ND1. Finally, we ask participants to review the requirements list, and to answer the following questions for each requirement:

- Is the requirement satisfied or not satisfied based on the information given in the diagram?
- How would the participant evaluate the security requirement: is it good, bad, unnecessary, immeasurable, unrealistic, etc.?

The questions above are asked in a conversational style with an open-ended fashion where participants are free to comment, explain, and elaborate in their answers. Since this article is based on a qualitative research method, pilot data from P1 and P2 can still be used in our full analysis of data.

Evaluation of approach

Our qualitative research methodology, called grounded theory, employs a different evaluation from quantitative approaches. Quantitative research often follows a positivist tradition in which phenomena in the world is comprised of measurable objects [33,34]. In this tradition, problems can be represented with variables that are explained through statistical relationships that support the repeatability of the results [34,35], e.g. Bayesian models of decision-making. Alternatively, qualitative research is prominent in the social sciences that follows a naturalist interpretivist tradition [34,36,37], wherein the researcher observes phenomenon while avoiding unintentional interference and manipulation [36]. Recall from section "Research approach" above that we were careful in our interview process to keep questions open-ended and we avoided individually evaluating experts for their performance. Qualitative methods are preferred when investigators are interested in discovering hypotheses and constructing new theories; whereas quantitative approaches are suited to testing hypotheses and evaluating theories.

These differences affect the determination of sample size. In quantitative-controlled user studies, sample size is determined by statistical power calculations, whereas in grounded theory, the sample size is determined by "saturation." The point of saturation "is the number of participants in which adding new participants is unexpected to surface new observations." This number depends on the phenomena being observed [38]: in our case, we have three artifact types, which restricts our observations to security analyst reflections on those artifacts. Atran *et al.* [39] estimate that a minimum of 10 participants is needed to show consensus, while Guest *et al.* argued that a sample size of 6 could be sufficient, if there is homogeneity among participants in the sample [39]. In our sample, we reached saturation after 8 participants, but we continued to recruit 3 more participants to test whether new observations would contradict our existing findings.

In the remainder of this section, we report the results from our empirical evaluation: the artifact assignment and inter-rater reliability.

Artifact assignment

Due to self-perceived inexperience by participants and time limitations, not every participant analyzed all artifacts in the three categories we described in section "Research approach". The average total interview time per participant to complete each interview was 29 minutes. Table 2 presents the participant assignment to conditions: the shaded cells show the category of artifacts that participants attempted; cells labeled with "X" indicate that the participant spent at least 15 minutes analyzing the artifact. Because participants have varying skills and expertise, some participants invested more time than others analyzing certain artifacts. The order in which the artifacts were presented to different participants was randomized and the time allowed to complete the interview was limited to 60 minutes. Thus, not all participants reviewed all artifacts. The "Sum" column in Table 2 presents the total number of participants who reviewed each artifact.

Agreement and inter-rater reliability

Two raters (the first and third authors) applied the coding frame from Table 1 to the transcripts of participant audio recordings. We measured inter-rater reliability using Cohen's Kappa, a statistic for measuring the proportion of agreement between two raters above, which might be expected by chance alone [39]. We calculated Kappa for each participant, which ranges between 0.51 and 0.77

Table 2. Participants' assignment by artifact

Artifact	Participants											Sum
	1	2	3	4	5	6	7	8	9	10	11	
Source code			X	X		X		X			X	8
Data flow		X	X	X		X	X	X				7
Network	X		X	X					X		X	7

with a median of 0.62. These values are considered moderate to substantial agreement [40]. The coding times were 19 and 8 hours for raters 1 and 2, respectively. Rater 1 spent more time documenting heuristics and developing the method. In addition to the above time, 6 hours were used for the resolution of disagreements between the two coders. Table 3 shows the breakdown of the total 2595 coded statements in our final dataset by code (including the pilot participants P1 and P2).

Decision-making patterns

We now present the decision-making patterns that ground the SA framework in the data. We use the acronyms introduced in Table 1 to express the patterns as a sequence of coded observations across the interview transcripts. Findings from this section are going to motivate the discussion, analysis, and impact on security analysis that is present in the remainder of this article.

The classic SA patterns

Endsley suggests that experts who assess risky situations engage in a process of perceiving information, comprehending the meaning of that information, and then projecting what might occur in the future. We call this pattern the "Classic SA" pattern, which proceeds from P→C→J→D where the "→" means the coded statement on the left-hand side appeared adjacent and before the coded statement on the right-hand side in the transcript. In addition to the Classic SA pattern, we searched for contiguous fragments of the Classic SA pattern in longer sequences, such as P→C→J, and C→J that indicate when a participant moves to higher levels of SA.

Table 4 presents the pattern name, number of occurrences (Frequency), and the accuracy (Accuracy), which is the ratio of actual, confirmed pattern instances among the total number of observations of the sequence (after removing false positives), and, finally, the list of participants who exhibited these patterns. We believe the pattern J→D is interesting because in combination with other patterns, we see variation fragments of the order appear. The results indicate that the J→D pattern only appears 31 times with 10% false positives. This observation suggests that projections and decisions, as well as other SA levels, can occur out of sequence, which motivated our search for the other pattern fragments shown in Table 4; all of these fragments are variations of the full Classic SA pattern (P→C→J→D). We observed that participants demonstrated the J→D pattern without the P→C pattern component, but this does not mean that participants did not perceive cues or comprehended those cues. Instead, participants may not be verbally reporting their perceptions and comprehension, or they may have automatized these stages of SA as part of their prior experience.

Table 3. Final dataset frequencies by code

Code	Total codes	Code	Total codes
Perception	250	Uncertain perception	82
Comprehension	498	Uncertain comprehension	180
Projection	215	Uncertain projection	13
Decision	367	Uncertain decision	25
Question	95	Probe	535
Background	47	Assumption	45
N/A	243		

Except for the first two patterns, a common feature among the patterns in Table 4 is the "skip" factor. Participants could skip a level of SA before reaching the next expected SA level. Because we coded participants' verbal responses, and participants may not have verbalized each level of cognition, our dataset may be missing the expressions of some levels. Another explanation for skipping levels is the level of expertise and exposure to the problem. If the participant has seen several examples of a certain problem, they may jump to their decisions immediately without providing explicit verbal analysis of the perceived cues, meanings, and possible consequences. The following is an example from P3's response to the SC artifact where they immediately projected an SQL attack without perceiving or comprehending a certain cue (we use brackets [] to explain the item of the artifact that the participant is speaking about):

> **{J}** this [speaking about the line of code that shows the unsanitized input] is just pure SQL injection here **{/J}**

By comparison, P11 articulated moving from perception to projection while describing the same attack scenario:

> **{P}** And thus, [speaking about the line of code that shows the unsanitized input], you use SQL query that explicitly say its inserting into the customer value **{/P}** **{J}** it may suffer from the SQL injection attack. **{/J}**

In contrast, the classic skip projection pattern from Table 4 describes how a participant moves from perception to comprehension but jumps to the decision phase without describing the projection.

The patterns (C→J) and (C→D) bypass the perception level, where participants move from comprehension to either a projection or a decision phase. Based on our analysis, it is not unusual for participants to begin verbalizing at the comprehension level. In this case, participants begin by describing the meaning of a cue without explicitly identifying the cue. Consider the following excerpt from the coded response of P9 when they were analyzing the Demilitarized Zone in the network artifact:

> **{C}**. . . people can access this part [speaking about the DMZ subnet in the network diagram] but it means de-militarized zone. **{/C}** **{J}**If these machines are hacked, they can't affect other inner parts **{/J}**

Table 4. Variations of classic SA patterns

Name	Pattern	Frequency	Accuracy[a] (%)	Participants
Classic w/o decision	P→C→J	4	100	P1, P3, P6
Projection-decision	J→D	31	90	All except P1
Classic skip projection	P→C→D	10	100	P1, P3, P4, P6, P11
Classic skip perception	C→J	55	81	All
Classic skip perception and projection	C→D	56	83	All except P2 and P5
Classic perception comprehension	P→C	60	81	All except P10
Classic perception comprehension followed by uncertainty	P→C→U*	7	74	P2, P3, P4, P7, P11

[a]Excluding false positives.

The pattern P→C in Table 4 reflects that participants move from the perception to the comprehension level, but without going immediately into projection or decision levels. We find these interesting because it shows that someone could move back and forth between perception and comprehension without moving higher to projection or decision. In fact, as the last pattern in the table shows, there were five (5/7 = 74%) instances where participants moved to uncertainty. We also found three more instances (100% accuracy) where participants start asking questions to resolve ambiguities (P→C→Q). This movement could indicate that participants found themselves "stuck" at comprehension where they could not proceed further, because they lacked the needed cues and understanding to envision what comes next or how to mitigate a threat.

The reverse SA patterns

In our dataset, we observed that SA patterns might occur in reverse order. This difference may be due to the participant using an inductive vs. deductive reasoning style. Up until now, we assumed that participants used a deductive reasoning style: they first report perceiving a cue, comprehending the meaning, and from this information, they deduce and report what may occur in the future (projection). In an inductive reasoning style, the participant verbalizes the possible consequences and from this information, they work backward by inducing the cues that led them to this conclusion. To accommodate the inductive reasoning style, we checked the dataset for patterns in the reverse direction of the classic SA pattern. Table 5 presents the reverse SA pattern names, their frequencies, accuracy, and participants who exhibited these patterns.

The following excerpt illustrates the reverse pattern exhibited by participant P6 who is analyzing the SC; the participant first reports their decision to prioritize a particular part of the diagram, followed by their understanding of this part and their perception of the part's character that led to the prioritization decision:

> {D}It's very important [speaking about using encryption for communication over the Internet] {/D} {C} you're sending the SSN over the Internet{/C} {P}it's [speaking about the SSN. . .] in plaintext. {/P}

Patterns of uncertainty and assumptions

Uncertainty plays an important role in security, as many security risks are probabilistic and participants must estimate the likelihood of particular events when forming projections. Moreover, analyst experience is likely to play a role in interpreting ambiguity in a specification and then deciding whether that ambiguity includes an interpretation that may lead to a security exploit. Table 6 presents the uncertainty patterns that we identified in the data. These patterns consist of statements coded with uncertainty (UP, UC, UJ, and UD) and assumptions {A}, questions {Q}, and decisions {D}. The total coded subset relevant to this discovery is comprised of 440 statements across all participants.

We categorized uncertainty into three categories:

- *Propagated Uncertainty* occurs in the first three patterns, wherein the uncertainty in perception or comprehension is propagated to a subsequent comprehension, projection or decision.
- *Hedged Uncertainty* occurs in all patterns where uncertainty leads to assumptions (e.g. U*→A), in which case the analyst bounds the uncertainty by interpreting an ambiguity and concluding this interpretation in the form of an assumption.
- *Uncertainty Transfer*, in which the analyst asks a question (e.g. U*→Q), to resolve uncertainty by seeking outside assistance.

With hedged uncertainty, 5 out of the 8 participants who made assumptions after their uncertain comprehension were able to make decisions. We found nine instances of hedged uncertainty leading to decisions, which may involve unstated assumptions. Finally, we observed that participants could move from a certain state to an uncertain one. In our dataset, we found participants transitioning to uncertain comprehension from perception (P→UC, 22 occurrences, 86% accuracy) or from comprehension (C→UC, 25 occurrences, 68% accuracy). Recall from section "The classic SA patterns" above how participants transitioned to uncertainty from the P→C pattern.

Patterns showing redundant states

In addition to the patterns we discussed in section "Decision-making patterns," we identified several patterns that appear to show the analyst is working harder to reach a decision. This includes patterns with accuracy rates above 60%: (C→C→C→C), (C→C→D), (P→C→C→J), (P→C→C→D), and (P→C→P→C). These patterns appeared 21, 26, 3, 5, and 12 times, respectively. The patterns show that participants are working harder to comprehend and interpret meanings to make more informed decisions. The patterns and corresponding text indicate that, the more detailed and thorough participants' comprehensions were, the better and clearer their future projections or decisions. This may explain why a participant needs more than one comprehension to reach the projection or decision levels. Moreover, there could be situations where complex security

Table 6. Uncertainty patterns

Pattern	Frequency	Accuracy (%)	Participants
UP→UC	8	100	P1, P3, P5, P6, P9
UC→UJ	2	100	P2, P5
UC→UD	2	100	P1, P4
UC→A	8	75	P1, P2, P3, P9, P11
UC→A→D	5	100	P1, P3, P9, P11
UC→Q	7	100	P2, P3, P4, P5, P7, P9
UP→A	5	60	P1, P3
UP→Q	3	67	P1, P3, P5
UC→D	9	67	P1, P2, P5, P6, P8, P9, P11

Table 5 Reverse SA patterns

Name	Pattern	Frequency	Accuracy (%)	Participants
Reverse SA w/ decision	D→J→C→P	None	None	None
Reverse SA w/o decision	J→C→P	1	100	P6
Reverse SA skip projection	D→C→P	3	67	P6, P9
Reverse SA no perception	J→C	35	67	All
Reverse SA no perception no projection	D→C	46	75	All

projections rely on multiple cues and comprehensions. Moreover, the comprehension level is where the analysis and interpretation begins, and projecting or forming a decision relies heavily on how well the analyst understands the vulnerability. For example, when an analyst comprehends the meaning of a firewall on the network, they consider different factors, which could lead them to verbalize more than one comprehension. Consider the following example as P3 was trying to analyze the network diagram ND2 against the first security requirement from the requirements list provided:

> {P}your firewall{/P} {C}which is your first point of entry to both DMZ traffic and intranet site traffic and also to your users{/C} {C} has all of these on separate subnets{/C} {D}the first rule here about stuff being unavailable comes down to whether this firewall is properly configured. {\D}

Participant P3 in the example above cannot reach a decision without comprehending two cues: (i) the firewall is the first point of entry to multiple network segments, and (ii) the firewall places the segments on different subnets. Therefore, this decision is dependent on a composition of multiple comprehensions, which explains the redundancy in the above pattern.

The SA path to security analysis

From our analysis results, we extended Endsley's SA model to account for uncertainty, the role of assumptions, and participant inquiry that results from uncertainty. Endlsey defines the stages of SA as they occur in the human mind, but since we are annotating participant articulations of those stages based on their verbal statements, there will be no guarantee that we will observe patterns in the data that will exactly reflect the classic or reverse SA workflow (P→C→J→D).

Hence, we decide to view SA levels as states where a security analyst could take different paths transitioning between the states. By our extended definition of SA, we open our analysis into other possibilities and combinations that would help understand security expert's decision-making process, and distinguish between experts and novices. We will elaborate more on this in the following sections.

Participants' expertise and the attacker model

We investigated whether more experienced participants would exhibit better SA and, thus, be able to form more confident decisions. Herein, we report our findings drawn from demographic data including participants' background and experience, and their experiences reported as remarks during their interview that we coded as {BG}. Next, we examine the role of expertise in forming more confident decisions. Finally, we link an expert's situation awareness with the attacker model by assessing how experts are achieving security decision based on impersonating an attacker.

Participants' background and expertise

Table 7 summarizes participant backgrounds (including pilot participants P1 and P2): the *P#* which is used consistently throughout this article; *Years* is the number of years of industry experience, including internships; *Security Areas* are the general topics that best describe their industry experience; *Research Focus* are the topics that best describe their research experience; and *Degree* is their highest degree earned, or in progress; Among the total 11, 4 participants (P1, P3, P4, P5) have extensive industry experience in security (4–15 years) with diverse concentrations.

P1, and P4 hold a PhD in security and specialize in systems and infrastructure. These two PhDs and P5 have teaching experience in which they taught advanced security courses. The remaining seven participants were all PhD students with research specialties in security. The PhD students had varying levels of experience, from a student who completed security courses, but who did not apply these lessons in practice beyond class projects, to students who had completed internships with a reputable company working on infrastructure security and log visualizations.

According to Endsley and Jones [17], an increase in experience may affect participants' ability to project future consequences and, hence, may lead to more confident decisions. In our study, we observe that participants with more industry experience were able to make more assumptions compared to those with less experience. For example, participants with more than 5+ years of industry experience made an average of seven assumptions, while participants with less than 5 years of experience made an average of one assumption. We coded statements with assumptions when the participant explicitly mentions that they are missing relevant details and that they have to assume or guess to complete their understanding.

Difference in artifacts presentation and notation could possibly affect situation awareness. Certain portions of an artifact were likely more unclear than others, so we may only expect to see assumptions when participants encountered less clear portions of the artifact. The pattern (UC→A→D) in Table 6 was observed for experts P1,

Table 7. Summary of participants' background

P#	Industry		Research	Degree
	Years	Security areas		
P1	5+	Network, systems, forensics, and more	Mobile computing, forensics, systems security	PhD
P2	<1	Security protocols, social networks	Global cyber threat	PhD[a] (fifth year)
P3	15+	Systems, networks, programming, and more	NA	B.S.
P4	5+	Systems, networks, architecture, and more	Security for real-time critical systems and architecture	PhD.
P5	10+	Software architecture, secure programming	Software architecture	M.S.
P6	0	NA	Cyber and system security	PhD[a] (fourth year)
P7	0	NA	Android security, malware, static analysis	PhD[a] (fourth year)
P8	1	Infrastructure security, log visualization	Security and privacy	PhD[a] (fifth year)
P9	0	NA	Security analysis, network traffic	PhD[a] (second year)
P10	0	NA	Anomaly detection	PhD[a] (first year)
P11	0	NA	Network traffic	PhD[a] (fourth year)

[a]PhD student, followed by year of matriculation in parentheses.

P3, and P9, when they analyzed the network artifact, and was observed for P11 when they analyzed the SC artifact. Participant P11 demonstrates advanced understanding when analyzing the SC artifact by reaching 24 decisions and this participant was the only one to make 2 assumptions in that artifact.

The attacker threat model

Experts' security analysis entails projecting future attack scenarios, and then deciding on how to mitigate them. This aligns very well in SA as we are already coding projections and decisions. In security analysis, projection and decision are closely related, because security analysts may be trained to think like an attacker and have an attack model in mind [12,41]. With an attack model in mind, the analyst decomposes a future attack scenario into multiple steps that exploit vulnerabilities. Under SA, we expect this decomposition to first appear as perceptions and comprehensions of the vulnerabilities, which then lead to the conclusion or projected exploitation, and finally a commensurate decision to mitigate the vulnerabilities. For example, Participant P3, notes: "what could I do since I am looking at this code to do bad stuff," which is their reflection on trying to walk through threat models that could be relevant to the code segment under review. P3 further stated: "it's critical if you're trying to design something secure to try and get into the mind of an attacker. If you can't think like an attacker, then you don't know how to defend against an attacker."

We analyzed our dataset to measure how often security analysts employed the attacker perspective. In our study, five participants (P1, P2, P6, P8, P10) demonstrated the need to think like an attacker as demonstrated by the word "attack" in their statements while referring to how an intruder would act.

Our results show 45 instances of attack words used where participants demonstrate knowledge of an attack; out of which only 29 instances describe an application of the attacker model where participants describe how the attack is taking place. The total 45 statements include instances where participants are explaining attacks that they knew about from their background, but without relating that knowledge to the artifact being analyzed. For example, the word attack could show up in a {BG} statement without a relevant SA pattern. For our analysis, we are interested in the 29 instances where participants are actually "thinking like an attacker" by demonstrating an attack scenario. Table 8 shows our results from this analysis: the participant number (P#) who described the attack scenario; the frequency (Freq.) that the term attack appears, the security artifact (Art.); and the relevant in-context patterns associated with the word—the SA code of the statement containing the attack word is highlighted in bolded text to show the position within the pattern. Each participant can exhibit multiple, separate instances of thinking like an attacker, which we separated by artifact and in-context pattern.

Among the 29 instances of the word "attack," we observe that most instances (25/29) occurred in the projection stage of SA. In less than half of the instances (12/29), the projection was observed after the interviewer probed the participant to explain why they were perceiving, comprehending, or projecting prior to describing the attack scenario (coded as Pro→**J**). Participants P2, P5, P7 are absent from Table 8, so they do not demonstrate the attacker model in their analysis.

Attack scenarios can be simple, meaning a single vulnerability is exploited to achieve an attacker's goal, or complex, meaning that multiple exploits are needed. In our results, we may observe and measure the complexity of attack scenarios as a series of different

Table 8. Participants' use of the term "attack"

P#	Freq.	Art.	In-context pattern
P1	5	ND1	P→C→C→Pro→**J**
		ND2	P→C→**J**→C
		ND2	D→D→Pro→C→C→**J**→C→C
		ND2	U→J→Pro→**UJ**→Pro→J
		ND2	Pro→UJ→Pro→**J**
P3	3	ND1	P→C→D→Pro→C→D→C→D→J→ D→D→Pro→**J**
		ND2	D→J→Pro→J→Pro→J→Pro→**J**→C
P4	2	ND2	D→C→C→**J**
		SC	D→C→Pro→**J**→Pro→C→C→P→C
P6	4	SC	**J**→D→**J**→J→C→C→J→Pro→C→C→Pro→P→J
		SC	C→C→**J**
		SC	D→**J** →D→D→J→Pro→C→P→J
P8	3	SC	C→Pro→**J**→Pro→J→D
		DFD	C→C→D→**J**→Pro
		DFD	C→J→**J**→C→C
P9	1	ND2	Pro→**J**→J→D→UP→D
P10	7	SC	D→Pro→**J**→J→**J**→D→C
		SC	**J**→Pro→Pro→J→**J**→D
		SC	**J**→J→Pro→**J**→J→**J**
P11	4	SC	P→**J**→J→D→**D**
		SC	C→C→**D**
		ND2	D→**C**→C→UC

SA stages need to demonstrate how an attack occurs within an artifact. For example, P9 projects a password brute force attack by looking at one item: requirement R7 on the list that reads: "Company X will require strong passwords (eight characters with complexity) for all user accounts." Based on the brute force projection, P9 decides that eight characters alone are not enough for a secure password policy. On the other hand, let us consider the attack pattern that P1 and P4 found in ND1: our entity analysis shows that in order to demonstrate the possible attack on the insecure network, both participants where analyzing multiple items in the ND1 diagram: allowed inbound ports on the router, the web server, the DNS controller, and the mail server. P1 further explained:

> {**J**} From an attacker that has no other entry point he is going to look at these three things [speaking about the 3 allowed inbound ports shown on the router], and if they didn't have any DNS server inside, there will be no reason to have port 53 open {**/J**}

Using SA patterns, we can compare participants' analysis when looking at the same entity (see our explanation of entity analysis in section "Research approach"). For example, in Table 8, participant P1 presents the pattern (P→C→**J**→C) in ND2 by first perceiving server names (entity code: NAME), such as Alpha, Lima, Bravo, etc. Participant P1 comprehends the server-naming scheme and subsequently projects that an attacker discovering these names alone cannot tell the role or function of the servers. Based on our entity analysis that links SA codes to these servers across participants, we found that participant P11 perceived the same naming scheme in their analysis (Q→P→C→UC→C), but they were unable to project based on the meaning of the scheme and thus were unable to see the attack scenario. Instead, P11 asks questions and experiences uncertain comprehension due to the meaning of the naming scheme whether the scheme has any relevance to network security. Unlike P1, participant P11 stops at comprehension and does not proceed to projection or decisions. This is an example of how the same cue could be interpreted differently by experts of different expertise levels.

Our SA attack model shows how we can use SA to detect a certain expertise skill: "thinking like an attacker." A conclusion that is based on the background data alone that is shown in Table 7 above, might indicate that participants: P1, P3, P4, and P5 are the more experts compared to the remaining participants in the table who could be treated as novices. This classification, which could be referred to as "industry classification," is based on participants' clearly combining years of practical industry experience along with academic degrees. However, this classification does not take into account the personal skills that a security analyst might acquire through their job or academic learning. Our attack threat model, on the other hand, help address this limitation by identifying the experts who demonstrate who can "think like an attacker." Table 8 shows that in addition to P1, P3, P4, who are already identified experts based on their industry experience, P6, P8, P9, P10, P11 can also demonstrate the skill of thinking like an attacker.

Going back to Table 8, we observe that except for P11's ND2 pattern, all participants had their "attack" keyword appearing in a projection or a decision statement, which resonates with the definition of our projection statements where a future attack is described, and our decision statements where mitigations to an attack is explained. By looking into the details of P11's pattern (D→**C**→C→UC), we observe how the participant is stuck at the comprehension level where they demonstrate a level of uncertainty.

Observations across the three artifact categories

The three categories of artifact—SC, DFD, and network diagrams—were chosen to vary specificity in system design and operation in order to surface variations in analyst performance. We now discuss those variations based on our SA results.

The source code

Eight participants were presented with the SC artifact, of whom seven agreed to analyze it. Six out of the seven participants identified at least two major concerns: the risk of SQL injection attack and of unencrypted user data. The remaining one participant, who was P10 by the way, could not spot the SQL injection vulnerability although he was reminded by the interviewer more than once to look at the artifact and provide any possible security concerns he might have, or if he has any further comments, etc.

The level of analysis and the proposed solutions varied in detail between the participants. While some were able to explain what languages to use and what libraries to call, some found it sufficient to explain that there are more secure measures that exist and good programmers should know about it. To investigate this more, we looked at the coded statements of participants; and compared participant P10 to others who were able to spot the vulnerabilities. For this specific source code artifact, P10 had only 4 perceptions compared to 12, 9, 13 perceptions for P6, P8, P11, respectively. However, P10 had 30 comprehension statements, which is the same as P11 who had more perceptions. When we read some of the statements, we found that P10 spent more time comprehending the 4 perceptions and deviated away from the intended attack to demonstrate other types of attacks that could occur such as phishing. Although Table 8 indicates that P10 can actually demonstrate thinking like an attacker, results from our entity analysis showed that P10 was demonstrating possible attacks other than the SQL injection attack.

The data flow diagram

We found 4/7 occurrences of the (UC→Q) pattern in the DFD, as participants report being confused about the chronological order of diagram entities. In addition, the DFD shows higher comprehension uncertainties (49 UC statements compared to 24 UC statements for source code). From the participant responses, we infer that all seven participants agree that the diagram lacks specific details needed for analysis. This result was expected when we chose the artifact: we deliberately chose the diagram showing fewer details to assess how ambiguity could affect the results. In our data, we observe two participants (P2, P5) responding differently to the ambiguity although they have perceived the same cue. Participant P2 states that they do not understand the role of the digital signature shown on the diagram (UC). In contrast, the participant P5 responds to the same entity by challenging the uncertainty with a perception and scaffolding their analysis with an assumption to reach a decision.

> {UC}Okay. So presumably I'm not sending my digital signature in the clear. It's an encrypted session, right? {/UC} {P} But again that doesn't really show that here{/P} {A}so if we assume that's an encrypted session and that I am not sharing my digital signature with somebody{/A} {D}then this is trusted{/D} {J}but if my machine's been compromised and someone has my digital signature they could potentially publish things as me, right? {/J}

In addition to the findings above, our frequency data shows that participants made less-decisions for this scenario (99) compared to 108 decisions for the source code and 160 decision statements for the network diagrams. Broken down by participants, participants P3, P4, and P5 who have higher industry experience were more hesitant to decide on the security of the DFD scenario making 3,2,4 decisions, respectively.

The network diagrams (ND1 and ND2)

The network artifacts illustrate how expertise areas and job role affect decision-making. Recall from section "The attacker threat model" how participant P1, and P11 reacted differently to the same perceived cue of the server-naming scheme. When we matched participant background information from Table 7 with their decision-making patterns, we observed that a job role, such as P1's hands-on experience in networking, might improve the participant's comprehension of cues and lead them to better decision-making.

Contrary to the SC artifact, where participants look at a code snippet showing one distinctive vulnerability: the SQL injection, network diagrams describe a composition of IT components (servers, routers, etc.) in which each component may have its own vulnerabilities. Thus, participants must view these vulnerabilities together to reach certain categories of decision. These interactions can be overwhelming for participants, if no structure is imposed on how they conduct their security analysis. We observed three modes of security analysis: unstructured, semi-structured, and structured, which we now discuss.

The unstructured mode

Participants were provided the least amount of structure when they were presented with the insecure network diagram (ND1) that had minimal cues, text, and legends. Every participant had a different starting point for their analysis although they are looking at the same entities in the artifact. Table 8 shows that P1 and P3 demonstrated an attacker threat for ND1, and the entity analysis shows that the two participants were looking at the same entities to demonstrate a possible attack (see section "The attacker threat model"). Participant P1 began their analysis from the firewall and its possible rules for open ports and participant P3 was more focused on the insecure layout of the DNS, e-mail, and web servers. Both participants

reached similar mitigation techniques, such as using a DMZ, and network segmentation in order to reduce the attack surface.

The semi-structured mode

The diagram ND2 has more legends and cues. The icons are distinguished by type of entity and the text and legends provide more detail, such as IP address, server name, OS type, etc. When participants analyzed ND2, they showed more structured analysis than they did with ND1. Contrary to ND1, all participants here, novices and experts, started at the same cue: network segmentation. They recognized the network segmentation of users, administration, management and DMZ, and explained the security advantages of such designs. The diagram in ND2 clearly shows the segmentation using legends and color codes that makes the network segmentation more obvious. However, some participants were not able to explain by the diagram alone some of the network design decisions such as the reason for having two separate DNS servers one of which is present in the DMZ. We will show next how structured analysis helped address this problem.

The structured mode

After presenting the diagram ND2 to the participants, we presented the security requirements checklist. We observed individual differences among experts and novices when assessing a single requirement and linking it to the diagram entities. In general, except for P2 and P5, participants who were presented with ND2 had started performing better compared to the two modes above. By better, we mean that participants were able to speak about certain items in the checklist which they missed to mention when they looked at the diagram without a structure. For each requirement on the checklist, analysts made an effort to connect each requirement to entities in the diagram. Table 9 below shows the results of mapping requirements to entities in the diagram by the participants who were presented with ND2. P2, and P5 are absent from the table as they have stated that they don't see how such matching could be achieved. None of the participants shown in Table 9 managed to map R3 (shown in Appendix), which is about "hardening" the network. P4 stated that the rule makes no sense, as it cannot be "qualified" nor "quantified." P3 commented with: "that's not uncommon for compliance to do that, to just state in very general terms a requirement, and then it's a little loose interpretation as to whether or not you've met that compliance or not." Highlighted cells in the table indicate that participants stated that dependencies exist among the highlighted requirement. P1 found the requirements R11 and R12 to be related. P1, P3, and P11 found agreed that R9 and R10 are related, but P11 failed to point out the entities on the diagram that map to the requirements.

Mapping the requirements–entity matching data in Table 9 to experience and background data in Table 7, we observe that P1, P3, P4 who had more industry experience than P9 and P11, were able to match more requirements on the list.

Using our entity analysis, we compared participants' responses across entities in diagram ND2. Our analysis results indicate that the requirements list helps both experts and novices: the experts' attention focused toward a specific security component and helped them reach better-informed decisions, and the novices became aware of a requirement and/or its security justification. Consider requirement R12 that requires a split DNS policy: expert participants P1, P3, P4, and P9 were able to map requirement R12 to the split DNS servers shown on the diagram and to state that the network satisfies the requirement, and they were also able to explain why such requirement is important from a security standpoint. Participants P1, P3, P4, P9 demonstrated the patterns: (P→P→UP→P→UP→D), (P→Q→Pro→D→J→J→J→A→J), (Q→C→C→C→J→J), (C→P→J→D→Pro→D→UC→C→A→C→C→J→C→D), respectively.

We investigated why P3 and P9 had longer patterns, and we found that P3 was demonstrating an attacker's attempt against the DNS server and how the split DNS increases the difficulty for attackers to break into the system. Towards the middle of participant P9's pattern, the participant exhibits uncertainty about why this requirement in needed for the system's security and thus they made an assumption in order better comprehend and project before reaching their final decision. Participant P11, was able to state that the requirement R12 is satisfied based on the diagram, but was unclear why a split DNS policy is needed. This is a good example of how introducing structure to security analysis, could help novices become aware of essential security requirements.

Table 9 suggests that participant P4 provided more entities among all participants. Going back to our interview notes, we found that P4 took an alternative and more highly structured approach to analysis by drawing a table on a blank piece of paper, listing the requirements numbers, and documenting how the requirement could be satisfied given the information shown on the diagram. During the interview process, P4 has shown more depth when analyzing the results and had confidence in their security analysis. We use the word depth here because P4 was able to refine requirements into specification levels and write down system specification and software configurations that are essential to satisfy the requirement, and this observation did not occur with any of the other participants.

Threats to validity

In this section, we address threats to construct, internal and external validity.

"Construct validity" is whether measures actually measure the construct of interest [42]. In our study, the construct of interest is SA, which is comprised of the four levels previously mentioned. One threat to construct validity is the definitions of the codes for each level in the coding frame are ambiguous and not mutually exclusive, such that the codes are inaccurately applied to the wrong statements (i.e., the perception code, if misapplied, may not be measuring instances of perception). To address this threat, we had two researchers (the first and third authors) meet to first discuss the coding frame before applying it to the dataset, after which we identified points of disagreement and reconciled these differences in a subsequent meeting. Recall from section "Evaluation of approach," we computed the inter-rater reliability statistic Cohen's Kappa that showed a moderate to high agreement. Unfortunately, we cannot know when participants are making implicit or unstated assumptions before reaching their decisions. Personality may be a co-factor that can effect whether or not participants make assumptions, since assumption making may be related to over-confidence.

"Internal validity" refers to whether the conclusions drawn from the data are valid [42]. Based on our coding of the data, we inferred several decision-making patterns in the data set that we report in section "Decision-making patterns." The completeness of the data threatens internal validity, because participants have unspoken perceptions, comprehension, etc. To address this threat, we employed probing questions to prompt participants to make explicit their SA levels, and we checked our observed patterns for accuracy across the dataset, i.e., how many instances of the pattern were consistent with

Table 9. Participants' requirements mapping to entities in ND2

R#[a]	P1	P3	P4	P9	P11
R1	Firewall-1	Firewall-1	Firewall-1	DMZ	
R2	Proxy (Squid)	Firewall-1, DNS-1	Proxy (Squid)		
R3					
R4	Proxy (Squid)		Proxy (Squid)		Snort1, Snort2, ArpWatch
R5		Windows DC	Firewall-1, Firewall-2, Exchange mail server		
R6	Firewall-1, Firewall-2	Exchange mail server	Exchange mail server, Mail Server on DMZ, Firewall-1	Exchange mail server	
R7		Windows DC	Exchange mail server		
R8	Firewall-2	Firewall-2	Firewall-1, Firewall-2		
R9	Syslog	Syslog	Nagios, ArpWatch	Syslog	
R10	Windows NTP	Windows NTP	Windows NTP		
R11	Snort-1, Snort-2, ArpWatch	Snort-1, Snort-2,	Snort-1, Snort-2, ArpWatch		
R12	DNS-1, DNS-2, DMZ	DNS-1, DNS-2, DMZ	DNS-1, DNS-2, Firewall-1, Firewall-2	DNS-1, DNS-2	
R13		ArpWatch	Snort-1, Snort-2, ArpWatch		
R14	Windows MRTG, Nagios	Syslog	Windows MRTG, Nagios		
R15		Firewall-2			

[a]Requirements' descriptions are listed in Appendix.

our definition of the pattern. This process led us to discover the reverse SA pattern reported in section "The Reverse SA patterns," which corresponds to differences between western deductive and eastern inductive reasoning styles previously studied in psychology [20,43,44].

The decision-making patterns are independent of the correctness of the security decisions. To evaluate whether the patterns lead to more correct or higher quality decisions or better performance, we need either: (i) an empirically validated "system metric" that measures how system specification elements contribute to a continuous or ordinate system security variable; or (ii) an empirically validated "psychometric," which measures the degree of agreement among multiple experts about the effect of each decision on an overall system security variable. We believe the first metric would preclude the need for human analysts, and the second metric must address several challenges, including: (i) people are rarely experts in more than one domain [45], and security is comprised of multiple sub-domains (networking, programming languages, operating systems, etc.); and (ii) knowledge is "a dominant source of variance in many human tasks" [45], which means small differences in security knowledge (e.g., awareness of emerging vulnerabilities) could lead to significant differences in decision evaluation, even among people who are commonly viewed as security experts. When adding these challenges to that of security specification, which includes that specifications can be ambiguous and incomplete, measuring the variability among experts requires a more principled approach to measurement than convenience samples and Likert scales. Hibshi *et al.* have introduced an approach to measure security adequacy based on multi-level modeling that uses multiple experts to review ambiguous and incomplete specifications [46]. This method may lead toward new psychometrics that address the challenges of expert decision-making in security, in which case we may discover relationships between decision-making patterns and improvements in security correctness.

"External validity" refers to the extent to which the results of this study can be generalized to other situations [42]. Two researchers have validated the SA coding frame, which increases its reliability and generalizability, and which can thus be reused by other investigators to probe the effect of new security analysis methods on analyst understanding. However, This study is based on grounded analysis, which limits generalizations to only this data set. While some might argue that our findings are thus too limited, qualitative research such as this frequently contributes to theory generation that supports follow-on controlled experiments. For example, we identified several prospects for future research that includes whether one can capture, encode and transfer expert assumptions to novices to facilitate transitioning novices from comprehension to projection and decision-making, or how can we improve perception to reduce uncertainty? In a following study [46], Hibshi *et al.* report experimental findings where experts have shown scenarios consisting of multiple factors that affect security decision-making. They capture security decisions from multiple experts in the form of ratings of individual scenario factors. The study design draws participant attention to individual factors to isolate the effect of their cue perception on attack projection to quantitatively measure the level of a security adequacy decision. Multi-level modeling is then used to analyze experts' consensus on the scenarios so the results can reliably be modeled.

Discussion

In this section, we discuss our results and the major takeaways from our research, possible future directions, and the impact on security requirements research. Researchers use grounded analysis to extract hypotheses from the data that could be extended and built upon in future studies [29,38]. We discussed in section "Evaluation of approach" above that, although our sample size is small, the quality of the data generated by our research is sufficient to draw new hypotheses that are valuable to the research community. We summarize our findings below in the form of hypotheses derived from our analysis results.

Security requirements exist in composition

Security experts apply requirements to a context while accounting for dependencies among these requirements. When we presented analysts with a checklist for the ND2 artifact, analysts were relating the checklist to the entities they observed in the diagram. In section "The network diagrams (ND1 and ND2)," we observe that the requirements checklist introduced a structured analysis for security assessment. For example, analysts had to relate the requirements to

two entities on the diagram: DNS1 and DNS2, in addition to the placement of these entities on the network, and how the user subnet is segmented from the management subnet and from the DMZ. Only an expert who understands how these pieces work together can conclude with confidence that a split DNS requirement is satisfied by the given diagram. Interestingly, participant P1 linked two requirements on the list together. While evaluating the R12 (split DNS), P1 states: "it does seem to be met at the diagram" and adds: "more importantly I think its superseded by the bullets up about Network Intrusion Detection Systems" referring to requirement R11 about network intrusion detection systems (see Appendix). P1 might have used the term "more importantly" to suggest that implementing a split DNS would be a waste of time without proper network monitoring and auditing. Thus, we derive the following hypotheses that identifying supersession can improve resource allocation:

H1: Identifying and removing superseded security requirements can increase security as resources are freed for re-allocation to mitigate other, unaddressed threats.

Our entity analysis shows that P1, P3, and P9 made another link between R9 and R10: logging and time synchronization. The following is an expert from P1's response describing the relationship (pattern C→J):

{C}if you implement logging, time synchronization is critical, {/C} {J}because otherwise the logs get interleaved and your one event ends up happening at different time stamps{/J}

Our results of the attacker threat models that we have shown in section "The attacker threat model" above, shows how experts need to relate pieces of a puzzle together in order to project the full attack. We believe that this is another reason why security requirements needed to mitigate the attack also exist as compositional "pieces of a larger puzzle" wherein some requirements complement each other, and others, such as logging requirements, have essential dependencies to be effective. Based on our findings, we propose the following hypothesis:

H2: Identifying dependencies among security requirements that address multiple threat models increases resiliency.

If true, the first two hypotheses indicate that experts cannot assess requirements independent of one another (particularly when using a simple checklist), and that supersession and dependencies can be used to increase overall security.

Effective cues improve security analysis

Throughout the paper, we discussed how certain analysts were able to perceive certain cues in the artifacts, comprehend them, and then, project and decide on mitigations, accordingly. However, we also showed cases where novice analysts were facing uncertainty during comprehension about a cue, e.g., trying to make sense of its meaning or its possible consequences. In section "The source code," we showed how one analyst, P10, did not even reach perception; P10 failed to perceive the cue that leads analysts to project the SQL injection attack.

In addition to measure where analysts struggled to move past perception and comprehension, we assessed the effect of improving notations and visual cues by comparing performance between the two network artifacts, ND1 and ND2 (see section "Observations across the three artifact categories"), and also by comparing the analysis results of the DFD artifact. Recall from Table 8 how only one participant P8, was able to demonstrate an attack on the diagram. In section "The data flow diagram," we showed how participants exhibited increased uncertainty analyzing the DFD artifact, which indicates how notational elements (or lack thereof) introduce ambiguity, which has a negative impact on analysis. In section "The network diagrams (ND1 and ND2)," we showed how less experienced participant P11 was able to confirm that the split DNS requirement is met by the diagram, although they do not understand its necessity, which is an example that good notation could help novices become more aware of essential requirements.

For the observations above, we hypothesize:

H3: Increasing cue validity, or the probability that analysts link cues to architectural categories, will reduce uncertainty and increase threat projection.

Ambiguity and resolution

Increasing ambiguity leads to different interpretations by experts looking at the same artifact, and in turn leads to different decisions regarding the appropriate requirements to mitigate the threat. In our study, we intentionally chose the ND1 with minimal cues and information displayed to study the role of ambiguity in decision-making. Consequently, participants interpreted a router icon differently, as a router or firewall. Figure 2 shows the different interpretations of the same entity by four participants, including their statements in order of articulation coded by the SA method. When the notation was improved in ND2, we observed a positive effect on P1, for example. After later seeing the firewall icon in diagram ND2, participant P1 returned to ND1 to correct their prior interpretation to conclude that the ND1 icon was a router.

We hypothesize that ambiguity affects the security analysis in the following way:

H4: An increase in ambiguity increases uncertainty in decisions, and thus reduces likelihood of threat mitigation.

Participants could not comprehend effectively if they did not perceive appropriate cues that lead to a comprehension, and that could explain having uncertainty patterns appear in our dataset (see section "Patterns of uncertainty and assumptions"), which leads an expert to transition to the ambiguity stage in Fig. 2. When analyzing the DFD artifact, for example, one participant attempted to think of all possible interpretations given the absence of specific details from the diagram. In the excerpt below, we show how participant P3 assumed that encryption existed:

{UC}that doesn't really show that here [speaking about encryption session for sending the digital signature] {/UC}, {A}so if we assume that's an encrypted session and that I am not sharing my digital signature with somebody{/A} {D}then this is trusted{/D}

For ambiguity resolution, we propose the hypothesis below:

H5: Increasing expertise increases the ability to make assumptions that resolve ambiguities, thus increasing the likelihood of threat mitigation.

In a few cases of uncertainty, assumptions helped participants resolve the ambiguity and reach their decisions. Those assumptions were not arbitrary; they were based on former experience and best practices adopted for network security that experts had been exposed to. The following coded excerpt that was taken from participant P1 and illustrates such an assumption:

{UP}I don't see an NTP server on this network{/UP} {C}but I know that Windows Domain Controller can act as NTP{/C}, {A}so I am going to assume that when they install it they'll

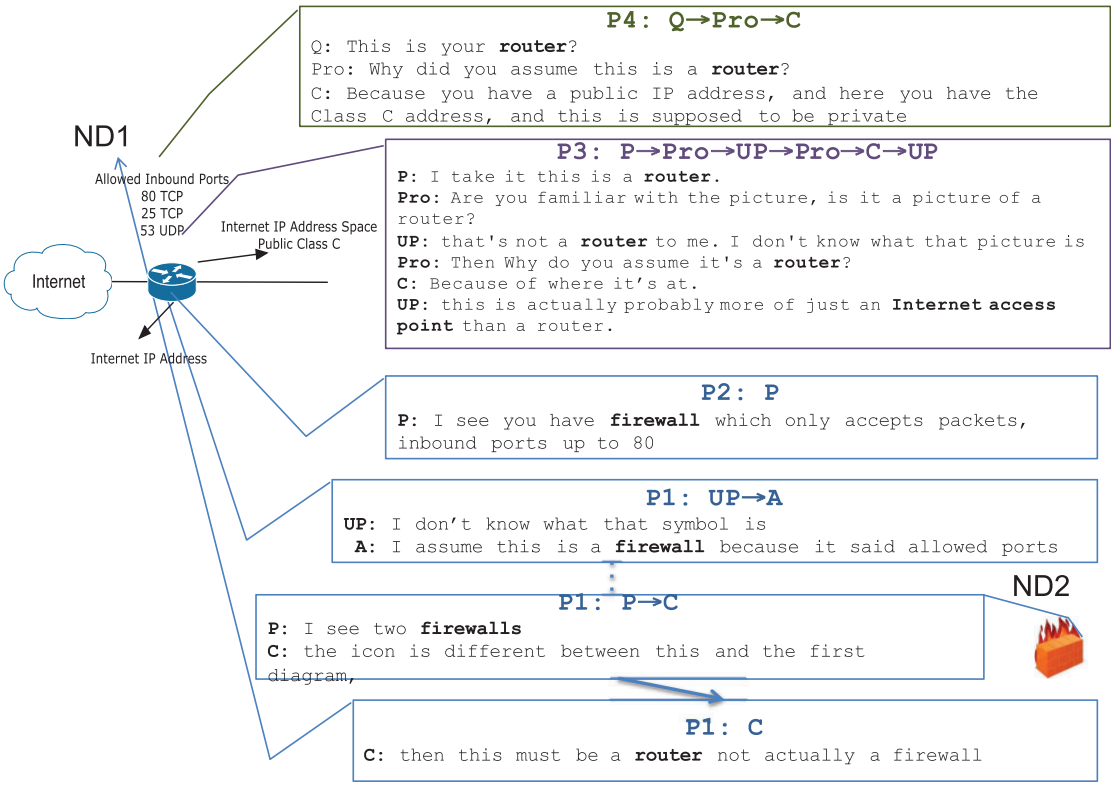

Figure 2. Participant perceptions of the router icon in diagram ND1.

probably leave that box checked because it's a default option {/A}. {D}I think that is probably happening here{/UD}

The above assumption is an example of a trust assumption first defined by Viega *et al.* [47] and then applied to security requirements by *Haley* et al. [14]. Trust assumptions describe desired behaviors and may be outside the control of the system designer. Based on the background-coded data {BG} (see Table 1 for a definition of this code), participant P1 has extensive hands-on experience in network security, which could explain why P1 was comfortable making assumptions about the system. The example above shows an interesting pattern (UP→C→A→UD). Although we did not observe the exact same pattern with other participants, we were able to observe the latter half of the pattern: A→UD as it occurred once for P5 and P11, and twice for P3 and P9. These participants reported significant experience in network security, so one would expect them to be more confident in reaching certain decisions with respect to network artifacts. However, we must not ignore the personality effect: an expert may hesitate to make confident decisions based on assumptions so they express a level of uncertainty with their decision to be more cautious.

Trust assumption reported by Haley *et al.* [13,14] help restrict the domain by narrowing the attention span of the analyst. In SA, a narrowed focus is beneficial for projection, but it can also lock-in the analyst and prevents them from perceiving alarming cues in the environment [17]. Moreover, incorrect assumptions about a system can lead to erroneous requirements specification [48]. Our work could be extended by distinguishing which assumptions are trust assumptions to distinguish the volatility of decisions that depend on assumptions about actors that are outside the system boundary. If those trust assumptions turn out to be untrue, then the security analysis that depends upon those assumptions should be revisited for possible inconsistencies.

Beresnevichien et al. introduced a methodology to support an organization's Chief Information Security Officers (CISO) in their decisions related to system security investment. The methodology elicits multi-attribute ratings of security preferences from decision makers and applies the utility functions to mathematically model the decision-making problem [49]. However, the approach does not capture the decision makers' uncertainties and the mathematical model is built with the assumption that the decision maker is confident about their choices.

In our dataset, we observed that experts were more likely to use assumptions to control uncertainty and to reach a decision. In future experiments, we could test if assumptions could provide another metric to distinguish between novices and experts. Being able to distinguish users based on expertise level could have an important impact on designing intelligent and interactive tools to help novice analysts cover more security scenarios in a problem description or specification.

Conclusions and future work

In this article, we present a new approach to assess security expertise and decision-making processes. Our contribution is: (i) a systematic method to apply the SA framework to distinguish security experts based on their differences in recognizing attack threat models; and (ii) new hypotheses regarding security decision-making and security requirements composition. We summarize our results to show traces across the SA levels in the form of patterns that could be used to distinguish experts from novices, and help identify factors impacting the security decision-making process. Our results suggest that security requirements checklists in which requirements are presented independently fail to capture the context of the attack and the composition of requirements. Composition can both increase and

decrease overall security assessments, because requirements interact to mitigate certain threats. We encourage researchers to take these factors into consideration when designing new security guidelines or decision-support systems.

Future work should further consider the level of uncertainty among security analysts while recognizing security experts will frequently be outliers in a diverse group. One approach to capture diverse expertise is through the use of user experiments, wherein experts are presented with short scenarios and asked to rate, prioritize, and decide on proper mitigations [46]. The study design will take into consideration capturing the uncertainty factor. In similar designs, the security scenarios will include cues to direct participants' attention to security mitigations to enact perception by asking the participant to rate the mitigation. Because security knowledge is distributed across sub-domains, such study designs should include a security knowledge post-test that is aligned with the study mitigations to measure whether participants can achieve the comprehension level necessary to project attacks and reliably rate the mitigation's contribution to security.

Based on the results of this work, we envision an adaptive security analysis system that can adapt to the training needs of a security trainee based on their perception and comprehension of cues. If a trainee fails to identify a cue, then the system could provide deeper training with further cues in order to help the trainee perceive vulnerabilities, comprehend its risk, project the impact, and decide on the proper mitigation. The SA application described in this paper helps to surface the cues that likely need to be supported in such a system. While experts may have little difficulties reaching projection and decision, novices may need additional information to aid them in reaching these higher levels. Identifying cues and testing their effectiveness, could lead to redesigning training artifacts in a way that makes the cues either more explicit (improve perception) or more meaningful (improve comprehension).

Finally, we believe that researchers can use the hypotheses produced from the SA study described herein to study how experts address security problems, to understand the wider spectrum of security experts that exist with different specialty domains in security, and most importantly, to provide solutions integrating security into our systems while accounting for all these decision-making factors.

Acknowledgments

We thank our study participants and Dr. Jennifer Cowley at SEI-CERT who consulted on Situation Awareness.

Funding

This work was supported by Army Research Office (Award #W911NF-09-1-0273); National Security Agency (Award #141333); and King Abdul-Aziz University.

References

1. HP. HP Top Cyber Security Risks Report. Hewlett-Packard Development Company, L.P., 2011.
2. OWASP. OWASP Top Ten Project - OWASP, 2014.
3. SANS. SANS 20 Critical Security Controls Solutions Directory, 2014.
4. NIST/ITL Special Publication (800), 2015.
5. Breaux TD, Baumer DL. Legally "reasonable" security requirements: A 10-year FTC retrospective. *Comput Secur* 2011;**30**:178–93.
6. Haley CB, Laney RC, Nuseibeh B, *et al*. Validating security requirements using structured toulmin-style argumentation. *Dep Comput Open Univ Milton Keynes UK Tech Rep* 2005;**4**:21.
7. Rittel HWJ, Webber MM. Wicked problems. *Man-Made Futur* 1974;**26**:272–80.
8. Dutoit AH, McCall R, Mistrík I, *et al*. (eds), Rationale management in software engineering: Concepts and techniques. *Rationale Management in Software Engineering*. Berlin, Heidelberg: Springer, 2006, 1–48.
9. Chung L. Dealing with security requirements during the development of information systems. In: Rolland C, Bodart F and Cauvet C (eds), *Advanced Information Systems Engineering*. Berlin, Heidelberg: Springer, 1993, 234–51.
10. McDermott J, Fox C. Using abuse case models for security requirements analysis. In: *Computer Security Applications Conference, 1999. (ACSAC '99) 15th Annual Proceedings*. Phoenix, AZ: IEEE, 1999, pp. 55–64.
11. Sindre G, Opdahl AL. Capturing security requirements through misuse cases. In: *Norsk Informatikkonferanse, NIK 2011*, Stavanger.
12. Van Lamsweerde A, Brohez S, De Landtsheer R, *et al*. From system goals to intruder anti-goals: attack generation and resolution for security requirements engineering. *Proc RHAS* 2003;**3**:49–56.
13. Haley CB, Laney RC, Moffett JD, *et al*. The effect of trust assumptions on the elaboration of security requirements. In: *Proceedings of 12th IEEE International Requirements Engineering Conference*. IEEE, 2004, pp. 102–11.
14. Haley CB, Laney RC, Moffett JD, *et al*. Using trust assumptions with security requirements. *Requir Eng* 2006;**11**:138–51.
15. Mellado D, Fernández-Medina E, Piattini M. A common criteria based security requirements engineering process for the development of secure information systems. *Comp Standards & Interfaces* 2007;**29**:244–53.
16. Endsley MR. Design and evaluation for situation awareness enhancement. In: *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*. Vol **32**. Thousand Oaks, CA: SAGE Publications, 1988, pp. 97–101.
17. Endsley MR, Jones DG. *Designing for Situation Awareness: An Approach to User-Centered Design*. Boca Raton, FL: Taylor & Francis, 2003.
18. Endsley MR. Toward a theory of situation awareness in dynamic systems. *Hum Factors J Hum Factors Ergon Soc* 1995;**37**:32–64.
19. Anderson JR. *Learning and Memory*. New York: John Wiley, 2000.
20. Anderson JA. Cognitive styles and multicultural populations. *J Teach Educ* 1988;**39**:2–9.
21. Bartlett FC, Burt C. Remembering: A study in experimental and social psychology. *Br J Educ Psychol* 1933;**3**:187–92.
22. Hintzman DL. "Schema Abstraction" in a multiple-trace memory model. *Psychol Rev* 1986;**93**:411–28.
23. Rao A, Hibshi H, Breaux TD, *et al*. Less is more?: Investigating the role of examples in security studies using analogical transfer. In: *Proceedings of the 2014 Symposium and Bootcamp on the Science of Security*. Raleigh, NC: ACM, 2014, pp.1–7.
24. Digioia G, Panzieri S. INFUSION: A system for situation and threat assessment in current and foreseen scenarios. In: *2012 IEEE International Multi-Disciplinary Conference on Cognitive Methods in Situation Awareness and Decision Support (CogSIMA)*. IEEE, 2012, pp. 316–23.
25. Feng Y-H, Teng T-H, Tan A-H. Modelling situation awareness for Context-aware Decision Support. *Expert Syst Appl* 2009;**36**:455–63.
26. Chen P-C, Liu P, Yen J, *et al*. Experience-based cyber situation recognition using relaxable logic patterns. In: *2012 IEEE International Multi-Disciplinary Conference on Cognitive Methods in Situation Awareness and Decision Support (CogSIMA)*. IEEE, 2012, pp. 243–50.
27. Schaefer KE, Billings DR, Hancock PA. Robots vs. machines: Identifying user perceptions and classifications. In: *2012 IEEE International Multi-Disciplinary Conference on Cognitive Methods in Situation Awareness and Decision Support (CogSIMA)* 2012, pp.138–41.
28. Jakobson G. Using federated adaptable multi-agent systems in achieving cyber attack tolerant missions. In: *2012 IEEE International Multi-Disciplinary Conference on Cognitive Methods in Situation Awareness and Decision Support (CogSIMA)* 2012, pp. 96–102.

29. Corbin J, Strauss A. *Basics of Qualitative Research: Techniques and Procedures for Developing Grounded Theory*. Thousand Oaks, CA: Sage, 2007.
30. Saldaña J. *The Coding Manual for Qualitative Researchers*. Thousand Oaks, CA: Sage, 2012.
31. May C. Applied Information Assurance, *Information Networking Institute*, [website], 2008, https://www.andrew.cmu.edu/course/14-761/ (24 June 2016, date last accessed).
32. Arasu A, Chaudhuri S, Ganjam K, *et al*. Incorporating String Transformations in Record Matching. In: *Proceedings of the 2008 ACM SIGMOD International Conference on Management of Data*. New York, NY, USA: ACM, 2008, pp.1231–34.
33. Glesne C, Peshkin A. *Becoming Qualitative Researchers: An Introduction*. New York, NY: Longman White Plains, 1992.
34. Golafshani N. Understanding reliability and validity in qualitative research. *The Qualitative Report* 2003;**8**:597–606.
35. Crocker L, Algina J. *Introduction to Classical and Modern Test Theory*. Orlando, FL: ERIC, 1986.
36. Patton MQ. *Qualitative Evaluation and Research Methods*. Thousand Oaks, CA: SAGE Publications, Inc, 1990.
37. Creswell JW. *Qualitative Inquiry and Research Design: Choosing among Five Approaches*. Thousand Oaks, CA: Sage, 2013.
38. Glaser BG, Strauss AL. *The Discovery of Grounded Theory: Strategies for Qualitative Research*. Piscataway, NJ: Transaction Publishers, 2009.
39. Atran S, Medin DL, Ross NO. The cultural mind: environmental decision making and cultural modeling within and across populations. *Psychol Rev* 2005;**112**:744.
40. Cohen J. Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit. *Psychol Bull* 1968;**70**:213.
41. Potter B, McGraw G. Software security testing. *Secur Priv IEEE* 2004;**2**:81–85.
42. Yin RK. *Case Study Research: Design and Methods*. Sage, 2009.
43. Choi I, Nisbett RE, Norenzayan A. Causal attribution across cultures: Variation and universality. *Psychol Bull* 1999;**125**:47.
44. Peng K, Nisbett RE. Culture, dialectics, and reasoning about contradiction. *Am Psychol* 1999;**54**:741.
45. Fletovich PJ, Prietula MJ, Ericsson KA. Studies of expertise from pscyhological perspectives. In: Ericsson KA, Charness N, Feltovich PJ *et al*. (eds), *The Cambridge Handbook of Expertise and Expert Performance*. Cambridge: IEEE, 2006.
46. Hibshi H, Breaux TD, Broomell SB. Assessment of risk perception in security requirements composition. In: *2015 IEEE 23rd International Requirements Engineering Conference (RE)*. New York, NY: ACM, 2015, pp.146–55.
47. Viega J, Kohno T, Potter B. Trust (and Mistrust) in secure applications. *Commun ACM* 2001;**44**:31–36.
48. Van Lamsweerde A, Letier E. From object orientation to goal orientation: A paradigm shift for requirements engineering. In: Wirsing M, Knapp A, Balsamo S (eds), *Radical Innovations of Software and Systems Engineering in the Future*. Berlin, Heidelberg: Springer, 2004, 325–40.
49. Beresnevichiene Y, Pym D, Shiu S. Decision support for systems security investment. In: *Network Operations and Management Symposium Workshops (NOMS Wksps), 2010 IEEE/IFIP*. IEEE, 2010, pp. 118–25.

Appendix

List of requirements used in artifact ND2

R1. Company X's network, with the exception of the publicly available services which will reside in a demilitarized zone (DMZ), will be unavailable for connections initiated from the Internet to Company X's network

R2. The employees of Company X will be required to use a web proxy server for connections to the World Wide Web.[WorldCat]

R3. Company X will harden and secure the services and operating systems of critical systems

R4. Company X will implement web content filtering and shall block inappropriate (pornographic) web sites

R5. Company X will implement a Windows domain, and will manage server and user system configurations through group policy centrally on the network

R6. Company X will implement an electronic mail relay, relaying mail from the Internet through a mail filter, which will filter spam and malware as mail enters Company X's network.

R7. Company X will require strong passwords (8 characters with complexity) for all user accounts.

R8. Company X will implement multiple networks (management, user, data center), and will implement strict access controls between each network.

R9. Company X will deploy system logging capabilities at all critical systems and will gather the logs centrally for review and response

R10. Company X will implement system time synchronization on the network for logging and auditing capabilities.

R11. Company X will implement multiple Intrusion Detection Systems (IDS) in multiple places on the network and shall audit regularly

a. File System Integrity IDS sensors shall be implemented
b. Network packet pattern matching IDS sensors shall be implemented.

R12. Company X shall implement split Domain Name System (DNS) services.

R13. Company X will monitor network traffic with packet sniffers.

R14. Company X will implement centralized system/service availability monitoring.

R15. Company X will administer all systems either interactively from the console or remotely from an isolated management network.