Assessment of Risk Perception in Security Requirements Composition

Hanan Hibshi[1,2], Travis D. Breaux[1]
Institute for Software Research, Carnegie Mellon Univeristy[1]
Pittsburgh, Pennsylvania, USA
College of Computing, King Abdul-Aziz University[2]
{hhibshi,breaux}@cs.cmu.edu

Stephen B. Broomell
Department of Social and Decision Science,
Carnegie Mellon University
Pittsburgh, Pennsylvania, USA
broomell@cmu.edu

***Abstract*—Security requirements analysis depends on how well-trained analysts perceive security risk, understand the impact of various vulnerabilities, and mitigate threats. When systems are composed of multiple machines, configurations, and software components that interact with each other, risk perception must account for the composition of security requirements. In this paper, we report on how changes to security requirements affect analysts risk perceptions and their decisions about how to modify the requirements to reach adequate security levels. We conducted two user surveys of 174 participants wherein participants assess security levels across 64 factorial vignettes. We analyzed the survey results using multi-level modeling to test for the effect of security requirements composition on participants' overall security adequacy ratings and on their ratings of individual requirements. We accompanied this analysis with grounded analysis of elicited requirements aimed at lowering the security risk. Our results suggest that requirements composition affects experts' adequacy ratings on security requirements. In addition, we identified three categories of requirements modifications, called refinements, replacements and reinforcements, and we measured how these categories compare with overall perceived security risk. Finally, we discuss the future impact of our work in security requirements assessment practice.**

***Index Terms* —user study; vignettes; factor surveys; security requirements; requirements elicitation.**

I. Introduction

Despite the abundance of well-documented security best practices, we continue to see security breaches that affect different organizations and industries. The 2013 OWASP Top 10 Application Security Risks report shows that attacks are occurring due to the exploitation of common, well-documented vulnerabilities, such as injection and cross-site scripting attacks [20]. In addition, security standards and best practices have long been available to help organizations improve security. For example, the U.S. National Institute of Standards and Technology (NIST) Special Publication (SP) 800 series describes best practice security requirements [19], and the Common Criteria describes a method to evaluate system security. In particular, the NIST SP 800-53 lists 256 security controls, which security analysts can apply in a checklist by deciding whether the control applies to their system. To make this decision, the analyst must reason over potentially millions of scenarios that account for various permutations of network type, services offered, threat type, etc. When requirements change by adding new components and features, these risk calculations must be updated. What is not known is how changes in threats and requirements affect the analyst's ability to perceive changes in risk and their ability to identify new, and reprioritize existing, security requirements.

In this paper, we report the results of two user surveys to study how changing threats and requirements affect the analysts' ability to perceive security risk and make corresponding decisions to prioritize security requirements. In designing these studies, we identified several properties of the problem of prioritizing security requirements:

- **Requirements Composition**: Security is a property of the composition of security requirements, i.e., it increases and decreases based on the addition and subtraction of specific requirements, while other requirements may have less effect. This is consistent with modern views of "defense in depth" [12] and layered security models, e.g. attack surfaces [17].
- **Requirements Ambiguity**: Abstract terms conceal multiple, possibly conflicting interpretations that can lead analysts to different decisions based on which interpretation they choose (e.g., "encrypted wireless network" includes both the insecure interpretation of WEP, and the presently secure interpretation of WPA2).
- **Requirements Completeness**: Because new threats and vulnerabilities continue to emerge, it is difficult to decide that the security requirements for a particular system are complete. Changes that affect completeness include new interpretations for existing security abstractions (e.g., WEP was viewed as an adequate security mitigation before the vulnerabilities were discovered), and the introduction of new system features through requirements composition.
- **Distributed Knowledge**: Knowledge of systems and their vulnerabilities is deep and distributed across multiple people (e.g., operating systems, networks, databases), which limit the ability of any one person to conclusively decide the priority of a system's security requirements. This distribution of knowledge further affects completeness, as one person is unlikely to be capable of evaluating the systems entire risk profile.

978-1-4673-6905-3/15 © 2015 IEEE

RE 2015, Ottawa, ON, Canada
Research Paper

Our survey design is intended to address these properties by eliciting risk perceptions from multiple analysts and targeting the mitigating effects of specific requirements to the threats they address. This approach allows us to isolate the effect of composition on security risk, and to address the limitations of differing levels of security expertise. In addition, our design asks analysts to report missing requirements. This step is aimed at improving completeness and reducing ambiguity.

The remaining paper is organized as follows: in Section II we review related work; in Section III, we present our survey design; in Section IV we present our multi-level modeling results; in Section V, we present our grounded analysis results; in Section VI, we present our threats to validity; in Section VII we present discussion, and we conclude in Section VIII.

II. Approach

We now introduce security vignettes, before describing the experimental design.

A. Vignettes Design

Our study is based on, factorial vignettes, which are scenarios comprised of discrete factors that contribute to human judgment. Researchers systematically manipulate the factors to understand their composite and individual effects on a decision [22, 26]. This reliable method is used by social and decision scientists and applied across psychology, sociology, and marketing, to name a few [2, 26]. Factorial vignettes are proven more effective to understanding decision making than direct questioning or single statement ratings that obscure the underlying contributions of different factors to the overall decision [1, 22, 26]. Factorial vignettes are presented in surveys using a basic template that contains multiple dimensions of the construct of interest. In our case, each dimension is a security requirement that influences the perceived level of security risk: some requirements increase risk, while others decrease risk. Figure 1 shows the template that we used in our study to create the vignettes: a vignette is a standard scenario generated by the template, wherein each variable name (starting with a $) is replaced by a level in the corresponding dimension.

> You are working on your laptop using **$NetworkType**. You are **$Transaction**. You are relying on a web browser to perform your task. The browser is already using **$Connection** for the session. To log in to the system and start your task, you will need to authenticate using a password that **$Password**. The system will **$Timer**.
>
> The **$Threat** is a serious security concern. Please answer the following questions with regards to mitigating this threat.

Fig. 1. The template used for vignette generation (fields with $ sign are replaced with values selected from Table 1)

In our study, each level corresponds to a requirement or system constraint variant, which is either a quality requirement (e.g., a "weak" vs. "strong" password) or more concrete interpretation of an otherwise ambiguous requirement (e.g., "unencrypted" vs. "encrypted" Wi-Fi). In Table I, we present the dimensions and levels to Fig. 1. Each level has a code (in parentheses) that we used to analyze and report our results. The Man-in-the-Middle `$Threat` occurs when an attacker intercepts the encrypted communication between two parties by decrypting the encryption. The Packet Sniffing `$Threat` is passive: the attacker eavesdrops on network packets to steal information without interacting with any parties, directly.

The choice of dimensions and levels in factorial vignettes is determined by the researcher's judgment based on the research questions. Our interest is evaluating the effect of changes in requirements composition and in threats where the composition spans a range of security knowledge, including network and application security, perceived sensitivity of information, and general "best practice" vs. threat-targeted mitigations. The dimensions that we chose are not the only dimensions that can be evaluated. In addition, the number of levels for each dimension is not the only number that exists.

In factorial vignette design, the space of all possible dimensions and levels is called the *factorial object universe* [22] and the *factorial object sample* is a sample across the universe used to instantiate the vignette template [22]. Sampling is random or systematic and the choice is based on prior theory, research, and reasoning [18]. Factorial sampling is used to eliminate unrealistic combinations of levels and to exclude scenarios that are likely to produce a predictable outcome [26]. Sampling from vignettes is more efficient than classic factorial designs, wherein all possible combinations of factors are tested [22].

We chose our initial scenario about logging into a remote e-mail service, because it crosses between novice and expert security knowledge, and this would allow us to measure the effect of security expertise on risk perception. We reviewed the universe and selected dimensions that had a sufficient number of levels to provide a rich space from which to sample; this includes network types and password complexity. Based on Table I, we have 32 (4×2×1×2×2) conditions per `$Threat` type.

TABLE I. Vignette Dimensions and Their Levels

Dimension	Level(s)
`$NetworkType`	`(EmpNetwork)` Your employer's network at your office
	`(PublicWIFI)` Public unencrypted Wi-Fi at a public area (restaurant, airport)
	`(VPNUnencrypted)` Your employer's VPN that you connected to through public unencrypted Wi-Fi
	`(VPNEncrypted)` Your employer's VPN that you connected to through public encrypted Wi-Fi
`$Transaction`	`(E)` Accessing your email account and replying to confidential emails
	`(F)` Performing a financial transaction using your credit card
`$Connection`	SSL
`$Password`	`(Weak)` A password that is at least 8 characters long
	`(Strong)` A password that is at least 16 characters and must include an uppercase and a lowercase letter, a symbol, and a number digit
`$Timer`	`(Yes)` Automatically log you off the session after 15 minutes of inactivity
	`(No)` Never time-out
`$Threat`	Man-in-the-Middle
	Packet-Sniffing

Our vignette selection is based on removing unrealistic and idiosyncratic scenarios. For example, the `$connection` dimension consists of one level, only, which is called a *blank dimension*. While we can evaluate unencrypted HTTP sessions in a scenario, the prevalence of knowledge about the high risk of unencrypted sessions suggests this level would predictably lead respondents to rate this requirement as inadequate to protect against the chosen threats. Blank dimensions are included in the vignettes, but not as statistical variables in the analysis, because they have no statistical effect to be measured. That said, blank dimensions are not to be eliminated, because their presence and absence affect how participants make decisions. In our case, removing SSL introduces an ambiguity: some participants may assume it exists, while others may assume it is absent. To control for this variability, we made this requirement explicit.

B. Survey Design and Research Questions

We designed our survey to answer three research questions:

RQ1. Does requirements composition affect risk perception in a security scenario to cause varied ratings of the security adequacy level, or can requirements be treated independently in a checklist style?

RQ2. Which security requirements in a security scenario contribute more weight to experts' security adequacy judgment?

RQ3. Would experts be able to detect ambiguities in a security scenario and provide modifications to improve the security adequacy ratings?

To answer these questions, we designed our survey instrument with three parts: the security vignettes, a security knowledge test, and a demographics test. In addition, each participant receives a consent form noting that participation is voluntary. We presented participants with the Man-in-the-Middle threat, where they answer all three parts of the survey. A week after taking the survey, participants are invited back for the Packet-Sniffing threat, where they do not repeat the security knowledge test or the demographic questions.

1) The Security Vignettes. In our study, each participant rates four vignettes to observe all the four network levels (see Table I). Since we have a total of 32 vignettes per threat, we have 8 possible combinations of the dimensions and, thus, each participant is randomly assigned to one of eight conditions, where they rate four vignettes (8×4 = 32 vignettes). Each condition randomly assigns the participant to a single level of the `$Transaction`, `$Password`, and `$Timer` dimensions (between-subjects effect), which are the same across all the four vignettes that the participant rates. The four vignettes differ by the `$NetworkType` dimension (within-subjects effect) and are presented in a randomized order.

For all four vignettes, a participant is asked to first rate the overall security level of the scenario within the context of the given threat. The rating levels are displayed in a random order from the following list:

- **Excessive** security measures that exceed the requirements to mitigate the threat
- **Inadequate** security measures that are not enough to mitigate the threat
- **Adequate** security measures that are enough to mitigate the threat

Next, we ask participants to rate the dimension levels based on the security requirement's ability to mitigate the given threat. This *mitigation rating* is applied to the `$NetworkType`, `$Connection`, `$Password` and `$Timer`, only, because they represent a mitigation that can be modified to improve security. Participants provide their rating on a 5-point Likert-scale, where point 1 is labeled "inadequate mitigation", point 3 is labeled "adequate mitigation" and point 5 is labeled "excessive mitigation." For each such dimension, we list the selected level for the vignette from Table I. These ratings are used to test which requirements (or factors) affect the overall security.

Participants are also given the opportunity to list additional security requirements that they believe contribute to increasing the security level to adequate. These are open-ended responses that we later analyzed by coding [23].

2) The Security Knowledge Test. Following the vignettes, participants are required to answer ten security knowledge questions. We selected these questions to cover user-level to administrator security knowledge, including cryptography, firewall rules, encryption, hashing, file permissions, and network security. The questions cover security concepts, and are intentionally inconvenient to search for on the Internet to reduce cheating. The responses are used to calculate a score that serves as a proxy experience metric.

3) Demographic Survey. Finally, participants answer questions about job experience and security training.

C. Pilot Study

The study was first piloted in three informal focus groups at Carnegie Mellon University. Attendees completed the survey; then critiqued the question prompts, response options, and vignette levels to eliminate unrealistic and idiosyncratic scenarios. The pilot stage was important to evaluate the security knowledge test questions, which were further refined, as our initial set of questions were lengthy and esoteric.

The survey was then piloted using the Amazon Mechanical Turk [1] (Mturk) crowdsourcing platform. We compensated participants with a $5 Amazon credit. Participants received one vignette from the 64-vignette sample. The pilot yielded feedback on response complexity and timing and led to the *mixed-methods* design, wherein we change the `$NetworkType` dimension for each subject 4 times. This design choice is effective in increasing the statistical power of the results [9]. In addition, the Mturk data shows a low mean for the security knowledge test (Mean = 4, Std. Dev.=1.87), and the responses to the open-ended questions were poor.

D. Deployment and Subject Recruitment

We recruited security experts using e-mail invitations to participate in our Man-in-the-Middle study (16 vignettes, where each participant sees 4 vignettes). The invitation was sent to security class mailing lists at Carnegie Mellon

[1] https://www.mturk.com/

University and North Carolina State University. We also sent invitations to security-research mailing lists at Carnegie Mellon University. We compensated participants a $10 Amazon gift card for participation. A week after taking this study, those participants were invited back to the Packet-Sniffing study (another 16 vignettes, where each participant sees 4 vignettes), and compensated with a second $10 Amazon gift card.

E. Analysis Approach

We now discuss our multi-level modeling and grounded analysis approach.

1) Analysis of Multi-level Models. Multi-level models are statistical regression models with parameters that account for multiple levels in datasets [9]. Our study design described in Section III.B supports both within and between subjects effects (mixed-effects). Thus, we treated the data as two studies based on the two levels of the `$Threat` dimension, which we assume the participant responses to the two threats are independent due to the week delay between surveys.

The quantitative dataset consists of one major outcome dependent variable: the `$OverallRating`, which is the security experts' judgment rating of the overall security level. This variable has three possible values -1, 0, or 1 that correspond to inadequate, adequate or excessive security, respectively. The fixed effects independent variables are the vignette dimensions: `$NetworkType`, `$Transaction`, `$Password`, `$Timer`, which we will refer to as requirements-mitigation variables. The random effect, independent variable is grouped by participant's `$ID`, because we have repeated measures for each subject who sees four levels of `$NetworkType`. We have four dependent mitigation-rating variables: `$NetworkRating`, `$Connection-Rating`, `$PasswordRating`, and `$TimerRating` that correspond to individual ratings of the `$NetworkType`, `$Connection`, `$Password`, and `$Timer` dimensions, respectively. Mitigation-rating variables are assigned an integer from 1-5.

We quantify experience using a `$Score` variable, which is an independent exploratory variable assigned an integer from 0-10 equal to the number of correct answers provided by the participant to the 10 security screening questions.

We analyze our data using multi-level modeling [9] to account for our mixed effect experiment design. We used R [21] and lme4 [3] as our tools to conduct the analysis. As described earlier, each participant rated all four levels of the `$NetworkType` dimension, while only rating one level of the remaining dimensions. Hence, our analysis simultaneously accounts for dependencies in the repeated measures, calculates the coefficients (weights) for each explanatory independent variable, and tests for interactions. We test the multi-level models' significance using the standard likelihood ratio test: we fit the regression model of interest; we fit a null model that excludes the independent variables used in the first model; we compute the likelihood ratio; and then, we report the chi-square, p-value, and degrees of freedom [9]. For fitted models that show statistical significance, we report the coefficient values from the regression model, which represents the dimension weight for predicting the dependent variable.

To determine sample size, we conducted a *priori* power analysis with *G*Power* [7] to test for the required sample size of repeated measures ANOVA. We estimated a sample size >96 per threat scenario for the recommended power level of 0.8 and a medium-sized effect [4].

2) The Grounded Analysis. We analyzed the mitigation requirements by first excluding non-mitigation responses. We then apply open coding [11] to code responses with short phrases (concept labels) and then group the phrases into six emergent categories: server, if the requirement is the responsibility of a web server, client, if the requirement is the responsibility of an application on the user's computer (e.g., a browser); encryption, if the requirement primarily concerns encrypting data or communications; private network, if the requirement suggests switching to a non-public network; attack detection and prevention, if the requirements is aimed at preventing and/or addressing certain attacks; and identity and authentication, if the requirement concerns verifying the identity of the user or their device.

After first cycle coding and categorization, we conducted a second-cycle coding [23], wherein we linked the categories to vignette dimensions and a *direction* as follows: a *refinement*, if the requirement refines the dimension by extending it's functionality; a *reinforcement*, if the requirement adds auxiliary security not directly related to the dimension; a *generalization*, if the requirement is more general than the dimension, but includes the dimension's mitigation; and a *replacement*, if the requirement replaces the dimension. For example, two requirements, multi-factor authentication and password expiry policy, are coded by the password dimension, yet the former is a *replacement*, because it replaces passwords with new functionality, and the latter is a *refinement*, because it extends passwords with expiration.

III. Multi-level Modeling Results

In this section we present the results of the multi-level modeling analysis.

A. Descriptive Statistics

A total 174 participants responded to the Man-in-the-Middle threat survey, of which, 116 returned to respond to the Packet-Sniffing survey. These sample sizes exceed what we estimated prior to conducting the study. The sample consists of 26% females and 73% males (1% unreported gender). The age groups sorted by dominance in the sample are 18-24 (63%), 25-34 (33%), and 35+ (3%). Within the sample there are 101 graduate students, 42 undergraduate students and 2 university professors.

TABLE II. Descriptive Statistics of the Rating Variables

	Man-in-the-Middle					**Packet-Sniffing**				
	Percentage*					***Percentage****				
	Adequacy Scale					***Adequacy Scale***				
	1		0		-1	1		0		-1
$OverallRating	0		53		42	0		92		1
Item Rating	***Adequacy Scale***					***Adequacy Scale***				
	5	**4**	**3**	**2**	**1**	**5**	**4**	**3**	**2**	**1**
$NetworkRating	1	9	37	21	32	2	7	36	22	33
$ConnectionRating	2	12	68	17	1	0	11	71	15	3
$PasswordRating	7	17	43	21	12	8	13	39	26	14
$TimeRating	2	11	29	17	41	4	12	27	21	36

*Percentages are calculated with respect to each threat study sample;
adequacy scale 5=Excessive, 3=Adequate, 1=Inadequate

The average number of participants per vignette is: 22 for the Man-In the-Middle threat, and 15 for the Packet-Sniffing threat; the number of participants is close but not equal across vignettes due to randomization. Table II presents descriptive statistics of participant ratings.

B. The Overall Rating

The $OverallRating variable is the major outcome dependent variable of interest, because this variable represents the experts' security rating of the scenario based on the composition of the requirements. Equation 1 is our main additive regression model with a random intercept grouped by participant ID. The additive model is a formula that defines the $OverallRating in terms of the intercept (α) and a series of components. Each component is multiplied by a coefficient (β) that represents the weight of that variable in the formula. This formula in Eq. 1 is simplified as it excludes the dummy (0/1) variable coding for the reader's convenience.

$$\$\text{OverallRating} = \alpha + \beta_{N}\$\text{NetworkType} + \beta_{Tran}\$\text{Transaction} + \beta_{P}\$\text{Password} + \beta_{Time}\$\text{Timer} + \epsilon \quad (1)$$

We will refer to the predictor explanatory variables: $NetworkType, $Transaction, $Password, and $Timer in this model as the *four predictors*. The β parameters in Eq. 1 represent the weight of each dimension in explaining the data. We tested the significance of the main effects in the additive model (Eq. 1); and then the interaction terms, which are the added terms generated by multiplication of the explanatory variables terms in the additive model. The indicator variables are dummy coded (0/1) to represent the dimension levels (see Table I). To compare the $OverallRating across vignettes we establish a *base level* for each variable that fixes the variables. The intercept (α) is the sample's mean outcome in the base case, which includes the following base levels:

- Employer network for the $NetworkType,
- Email for $Transaction,
- Strong password for $Password and,
- No timer for $Timer.

For the Man-in-the-Middle threat, we found a significant contribution of the four predictors for predicting the $OverallRating ($\chi^2$ (6)= 142.2, $p < 0.001$) but failed to find a significant contribution from the interaction terms (χ^2 (11)= 4.8, p = 0.94). For the Packet-Sniffing threat, the $OverallRating is also affected by the same four predictor variables with a significant value over the null model (χ^2 (6)= 20.4, p=0.002). We also did not see any significance for the interaction model (χ^2 (11)= 6.6, p = 0.83). These results suggest that the four dimensions $NetworkType, $Transaction, $Password, and $Timer are good predictors that explain change in the expert's overall rating. However, it is important to note here that the dataset from the Packet-Sniffing threat is less predictive in explaining the $OverallRating variable due to the violation of the normality assumptions. This is due to the unforeseen effect of no participant choosing the excessive rating in vignettes with this threat type (see Table II), which reduced the response levels from three to two.

Table III shows the assigned coefficient weights (labeled by β in the table) along with standard errors and significance levels for the two threat datasets. These weights represent the amount of change in rating caused by the corresponding change in predictor variable level. From the table, we conjecture that the Man-in-the-Middle threat has a significant intercept of –0.24, which indicates that at the base case (employer's network, email transaction, strong password, and no timer), the mean of the $OverallRating is lower than adequate (Recall from Section III, adequate= 0). Since, NetworkType is the only dimension showing significance in the table, we further interpret the intercept to indicate the mean adequacy level in the case of the employer's network. Interestingly, the public Wi-Fi network and the VPN over unencrypted network significantly decreased the overall rating from the base level employers network. Another interesting observation in Table III is that the VPN over encrypted network significantly increase the overall rating in the Packet-Sniffing threat scenario, while this has no effect in the Man-in-the-Middle threat. This result is expected from security experts who understand the difference among the two threats: encryption is a reasonable protection against Packet-Sniffing as attackers would not benefit from sniffing encrypted packets, but encryption alone is not enough to mitigate Man-in-the-Middle wherein attackers intercept and decrypt encrypted communication.

The significance levels in Table III indicate $NetworkType is the only dimension that had an effect on experts' $OverallRating of the security scenario. This does not mean the other dimensions had no effect on expert judgment. These estimates imply that the network type had the most influence (weight) on judgments of overall rating and the importance of each network type depends on the type of $threat.

TABLE III. RESULTS OF REGRESSION FOR THE $OVERALLRATING VARIABLE

Variable-level	Man-in-the-Middle	Packet Sniffing
	β (Std. Error)	*β (Std. Error)*
Intercept	–0.24 (0.07)***	0.10(0.06)
Network-PublicWIFI	–0.50(0.05)***	–0.03(0.05)
Network-VPNEncrypted	0.03(0.05)	0.10(0.03)**
Network-VPNUnencrypted	–0.24(0.05)***	–0.04(0.04)
Transaction-F	–0.03 (0.06)	–0.02(0.04)
Password-weak	0.06(0.07)	–0.05(0.05)
Timer-Yes	0.14(0.08)	0.03(0.06)

*p≤.05 **p≤.01 ***p≤.001, with standard errors in parentheses

C. The Security Requirements Effect

In this section, we further examine the effect of each requirement in the security scenario by analyzing participants' 5-point Likert-scale ratings of the specific mitigations. To do this analysis, we use the same regression formula in Eq. 1, but replace the $OverallRating outcome variable with $NetworkRating, $ConnectionRating, $PasswordRating, or $TimerRating. We now discuss the requirements effects.

1) The Network Effect. The $NetworkRating is a measure of the participants adequacy rating of the network in the scenario in order to get more insight into how they formed their $OverallRating of the scenario. We found a significant contribution of the four predictors for predicting the $NetworkRating. This significant result applies to both threat scenarios: Man-in-the-Middle (χ^2 (6)= 322.1, p<0.001), and

Packet-Sniffing (χ^2 (6)= 209, p<0.001). As with $OverallRating, we did not find any added significance from the interaction terms for both threats: Man-in-the-Middle (χ^2 (11)= 6.4, p=0.84), and Packet-Sniffing (χ^2 (11)= 6.3, p=0.85).

TABLE IV. RESULTS OF REGRESSION FOR THE $NETWORKRATING VARIABLE

Variable-level	Man-in-the-Middle	Packet Sniffing
	β (Std. Error)	*β (Std. Error)*
Intercept	2.70(0.10)***	2.43 (0.14)***
Network-PublicWIFI	–1.28(0.08)***	–1.13(0.10)***
Network-VPNEncrypted	0.35(0.08)***	0.47(0.10)***
Network-VPNUnencrypted	–0.35(0.08)***	–0.18(0.10)
Transaction-F	–0.14(0.08)	–0.08(0.10)
Password-weak	0.07(0.09)	0.06(0.13)
Timer-Yes	–0.06(0.11)	0.05(0.14)

*p≤.05 **p≤.01 ***p≤.001, with standard errors in parentheses

Table IV shows the detailed results of the regression model for the $NetworkRating outcome variable. From the intercept value, we conjecture that participants' rated the base case (employer's network, email transaction, strong password, and no timer) slightly lower than adequate (recall from Section III, adequate = 3). The table also shows how the network type has a significant effect on the $NetworkRating variable. In both threat scenarios, changing from the employer's network to the public Wi-Fi network decreased the rating by more than one point. On the other hand, the VPN over encrypted Wi-Fi significantly increased the $NetworkRating adequacy level over the employer's network. For the Packet-Sniffing threat, the VPN over unencrypted network did not have an effect on the network rating for that threat. This means that participants view the VPN over unencrypted Wi-Fi and the employer's network to be at the same security adequacy level.

Another observation from the table is the absence of effect for the other requirements on the $NetworkRating adequacy. There are two possible explanations for this result: 1) when participants are rating the network, they isolate it from all other requirements and they only focus on looking at the network type, and/or 2) participants are assigning a higher priority to the $NetworkType so it acts as the deciding factor and it supersedes other requirements in the scenario.

2) The SSL Connection Effect. We found slight statistically significant contribution of the four predictors predicting the $ConnectionRating adequacy level for the Man-in-the-Middle threat (χ^2 (6)= 15.1, p=0.02), but no significant contribution in the Packet-Sniffing dataset (χ^2 (6)= 5.8, p=0.5). When we further examined the regression model of the Man-in-the-Middle dataset, we found significance only for the intercept (α=2.9, SE=0.10, p<0.001) and the public Wi-Fi network (β = –0.10, SE=0.05, p=0.03). This means that the mean for the $ConnectionRating in the base case is around adequate, while it slightly drops when the network changes from employer's network to a public Wi-Fi. One possible interpretation of these results could be that the presence of SSL in the scenario is crucial and that's why the mean is around adequate, but the adequacy rating does not significantly change with the change of other requirements except if the change is to an extremely low level of security such as Public Wi-Fi.

3) The Password Strength Effect. The $PasswordRating is a measure of the participants adequacy rating of the password strength in the scenario.

The four-predictor model significantly increases model fit of $PasswordRating over the null model. This is present in both threat scenarios: Man-in-the-Middle (χ^2 (6)= 37.6, p<0.001), and Packet-Sniffing (χ^2 (6)= 38.6, p<0.001). Similar to the above outcome rating variables, the interaction terms do not significantly increase the model fit for the Man-in-the-Middle threat (χ^2 (11)=11.7, p=0.38). Although the Packet-Sniffing threat showed a significant effect (χ^2 (11)= 22.5, p=0.02), the coefficients did not show significant p-values for the interaction terms, which may indicate that the added significance was distributed across the terms.

Table V shows the details of the regression model for the $PasswordRating variable. In both scenarios, the intercept at the base case where the password is strong shows significant adequate ratings, that drops significantly when the network changes from employer's network (base case) to public Wi-Fi. Changing the password strength from strong to weak also drops the password adequacy rating in both threat scenarios.

TABLE V. RESULTS OF REGRESSION FOR THE $PASSWORDRATING VARIABLE

Variable-level	Man-in-the-Middle	Packet Sniffing
	β (Std. Error)	*β (Std. Error)*
Intercept	3.33(0.16)***	3.16(0.22)***
Network-PublicWIFI	–0.15(0.05)***	–0.18(0.05)***
Network-VPNEncrypted	–0.01(0.05)	0.06(0.05)
Network-VPNUnencrypted	–0.05(0.04)	–0.06(0.05)
Transaction-F	–0.05(0.14)	–0.14(0.19)
Password-weak	–0.76(0.17)***	–0.68(0.23)**
Timer-Yes	–0.10(0.20)	0.15(0.26)

*p≤.05 **p≤.01 ***p≤.001, with standard errors in parentheses

4) The Auto-logoff Timer Effect. The $TimerRating is a measure of the participants adequacy rating of the auto-logoff timer in the scenario.

The four-predictor model significantly increases model fit of $TimerRating over the null model. This is present in both threat scenarios: Man-in-the-Middle (χ^2 (6)= 54.9, p<0.001), and Packet-Sniffing (χ^2 (6)= 49.2, p<0.001). Similar to the above outcome rating variables, the interaction terms do not significantly increase the model fit for the Man-in-the-Middle threat (χ^2 (11)=17.4, p=0.09), or the Packet-Sniffing threat (χ^2 (11)=12.9, p=0.30).

Table VI shows the details of the regression model for the $TimerRating variable. Note that the intercept shows a low mean that is close to inadequate (recall from Section III: inadequate = 1) which is expected since the base level has no auto log-off timer. In the presence of the Man-in-the-Middle threat, the $NetworkType, $Password, and $Timer dimensions have a significant impact on participants' $TimerRating. The public Wi-Fi, VPN over unencrypted Wi-Fi, decreased the adequacy level of the $TimerRating variable, while turning the auto logoff timer on had significantly increased the adequacy

level of the $TimerRating. In the case of the Packet-Sniffing threat, the network type did not have a significant impact on predicting the $TimerRating, but the presence of the timer in the scenario shows a significant increase in the $TimerRating compared to the base case where no timer is involved.

TABLE VI. RESULTS OF REGRESSION FOR THE $TIMERRATING VARIABLE

Variable-level	Man-in-the-Middle	Packet Sniffing
	β (Std. Error)	*β (Std. Error)*
Intercept	1.79(0.17)***	1.51(0.22)***
Network-PublicWIFI	–0.18(0.05)***	–0.08(0.05)
Network-VPNEncrypted	0.0(0.05)	0.06(0.05)
Network-VPNUnencrypted	–0.12(0.05)**	0.07(0.05)
Transaction-F	–0.25(0.15)	–0.22(0.18)
Password-weak	0.60(0.18)***	0.82(0.22)***
Timer-Yes	1.18(0.21)***	1.60(0.25)***

*p≤.05 **p≤.01 ***p≤.001, with standard errors in parentheses

It is strange and unexpected that the weak password is showing a significant increase in the timer adequacy rating in both scenarios. It is possible that this is a Type I error (i.e., the password didn't actually play a role in the decision and this effect is only random) or is due to an interaction effect between password and the other predictor variables. When we examined the coefficients of the interaction model, we observed that the weak password significantly interacts with other variables such as public Wi-Fi and VPN over unencrypted Wi-Fi, which makes us lean more towards the interaction explanation although the data does not show evidence of interaction.

D. *The Experience Effect*

The $Score variable is our indicator variable for experience, as it represents participants score (out of 10) on the security test. Scored responses to our knowledge test presented a minimum score of 1 and a maximum of 10, with a mean 5.2 and a median of 5.

We added the experience predictor variable ($Score) to Eq. 1 and compared the new model to the four-predictor model in Eq. 1. for both threat types. The new model with the experience indicator ($Score) did not significantly improve the prediction of the overall variable compared to the original model with the four predictors alone. We repeated the same comparison for all the four mitigation-rating variables: $NetworkRating, $ConnectionRating, $PasswordRating, and $TimerRating. Except for the $PasswordRating and the $TimerRating in the Man-in-the-Middle threat, the ($Score) variable did not significantly improve the prediction of the ratings variables.

In the presence of the Man-in-the-Middle threat, adding the experience indicator ($Score) to the four predictor model improved the prediction of the $PasswordRating ($\chi^2$ (1)=1.8, p<0.001). The coefficient weight in the model shows that the $PasswordRating significantly decrease by –0.15 as the experience indicator ($Score) increases. Similarly, adding ($Score) to the four predictors model improved the prediction of the $TimerRating ($\chi^2$ (1)=8.2, p=0.004). The coefficient weight in the model shows that the $TimerRating significantly decrease by –0.11 as the experience indicator ($Score) increases. In other words, more knowledgeable participants (with higher $Score) tend to act more conservative when rating the adequacy level of the password and timer mitigations.

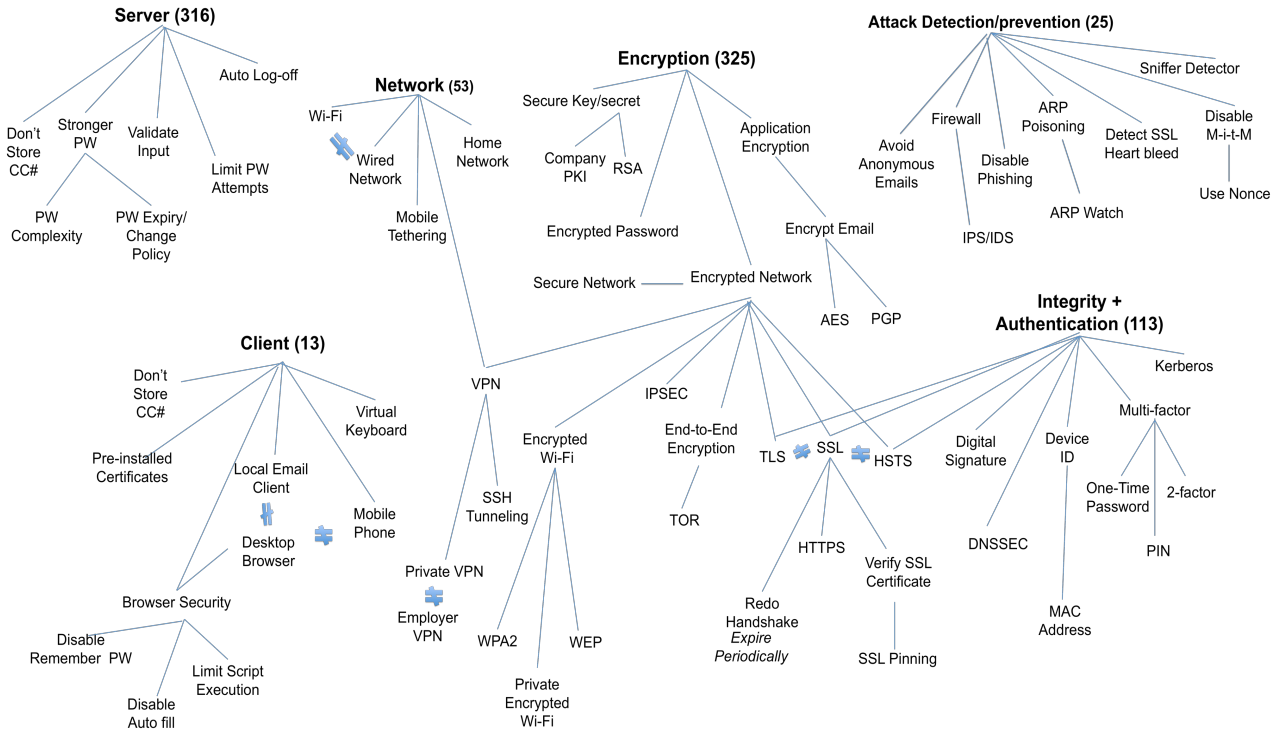

Fig. 2. The elicited requirements and their categories (numbers in parentheses correspond to number of statements)

IV. Grounded Analysis of Elicited Requirements

We elicited 905 mitigations from 108 participants: 540 for `Man-in-the-Middle` (104 participants) and 365 for `Packet-Sniffing` (64 participants). We organized the mitigations into 6 categories (see Section III). Figure 2 shows all 6 categories with mitigation concepts under each category. We analyzed elicited mitigations in response to the network effect, because our statistical results suggest that the `$NetworkType` has the most influence on participants' judgments. Table VII shows for each `$NetworkType`, the number of mitigations provided by participants (*Mits*.), the number of respondents providing these mitigations (*Resp*.), and total mitigations. Table VIII shows the number of refinements (*Refine*.), which are elaborations on an existing security requirement in the vignette (e.g., SSL, VPN); reinforcements (*Reinf*.), which describe auxiliary or new security functionality intended to complement existing requirements; replacements (*Repl*.), which describe a requirement to supplant an existing requirement (e.g., WPA2 supplants WEP); and generalizations (*Gen*.), which describe more abstract requirements (e.g., secure network v. VPN).

TABLE VII. Number of Mitigation Requirments by Threat and Network Type

`$NetworkType`	Man-in-the-Middle		Packet Sniffing		Total
	Mits.	*Resp.*	*Mits.*	*Resp.*	*Mits.*
Employer's Network	129	73	100	51	229
Public Wi-Fi	162	82	110	57	272
VPN over Unencrypted Wi-Fi	135	73	79	47	214
VPN over Encrypted Wi-Fi	114	73	76	42	190

TABLE VIII. Refinements, Reinforcements, Replacements, and Generalizations Requirements by Network Type

`$NetworkType`	*Refine.*	*Reinf.*	*Repl.*	*Gen.*	*Total*
Employer's Network	107	41	63	18	229
Public Wi-Fi	88	33	122	29	272
VPN over Unencrypted Wi-Fi	91	23	78	22	214
VPN over Encrypted Wi-Fi	101	23	57	9	190
Total	387	120	320	78	905

In Table VII, the weakest network type Public Wi-Fi has the highest number of mitigations for both threat types. Notably, Table VIII includes 155 auto-log off timer mitigations suggested by participants who observed no auto logoff timer in the vignette, and 107 complex-password mitigations suggested by participants who observed a weak password in the vignette. After removing such refinements that we expected to see in the lower security dimension levels, we found 125 refinements remaining. We now highlight some of the findings.

Several refinements served to remove ambiguity. For example, we found 51 mitigations that refine SSL, such as requiring updates or patching the *heart bleed* vulnerability [25]. One participant suggested using WPA2 encrypted Wi-Fi, because the Wi-Fi encryption was unspecified. Two participants stressed that VPN over encrypted network should use a reliably strong encryption.

Among reinforcements, we found 25 mitigations proposing *attack detection / prevention* techniques (see Fig.2), 24 mitigations adding email encryption under the email transaction condition, and 8 requirements to add browser security and pre-installed SSL certificates, among others. Some reinforcements were inspired by the vignette: four mitigations against man-in-the-middle attacks, four against packet sniffing, and two against email phishing attacks.

Replacement mitigations aim to replace a less secure requirement or constraint with a more secure alternative. We found 95 mitigations to replace the password with multifactor authentication. We also found 21 mitigations to replace SSL with TLS or HSTS, which is a recent security proposal receiving more attention [6, 16].

V. Threats to Validity

Internal validity is the degree to which a causal relationship can be inferred between the independent predictor variables and the outcome dependent variables [24]. In our study, we randomized the assignment to conditions and the order of the four vignettes shown to each participant. We also randomize the order of the 3 adequacy ratings in the overall security-rating question, and we mask the numerical values for these ratings from participants. To address the threats of learning and fatigue effects, we estimate a 20 minutes average time for each threat survey, and we maintain a time space of a week minimum between threat conditions. We did not randomize the threat scenario order, but we mitigated the effect of this decision by treating the two threats as separate datasets during analysis.

External validity concerns how well our results generalize to the population and situations outside the sample used in the study [24]. Our target population is security experts and we targeted participants by recruiting from senior and graduate level security classes. Furthermore, we conducted a security knowledge test to measure their expertise. One possible sample bias is that our sample was drawn from two U.S. Universities.

Construct validity concerns how well the measurements we take correspond to the construct of interest [24]. To ensure that participants have a shared understanding of the ratings, we provided one-sentence definitions for each rating level. Prior to choosing the three levels' labels, we tested 15 terms in an online survey of over 300 participants to find those terms that predictably associate with increased and decreased security. For the experience indicator (`$Score`) variable, it is important to note that this is the first time such test is used, and we would need more future studies to test its validity.

Reducing sources of variation in a study helps to increase *power*. We used different ways to improve power. For example, we instrumented a mixed-models design that combines within-subjects and between-subjects effects. We also analyze our data with multi-level regression modeling which limits the biased covariance estimates by assigning a random intercept for each subject [9]. In Section III, we discuss how we estimated the sample size needed for our study, and our final sample size is 81% higher than the estimated size.

VI. Discussion and Future Work

Our multilevel modeling results and grounded analysis suggest that risk perception varies with how requirements are composed, which addresses RQ1. These results also address

RQ2, because the dimensions/levels indicate the weighted contributions to the security adequacy ratings, which we now discuss.

We observed composition across the participants' `$PasswordRating`, `$TimerRating`, and `$ConnectionRating` and from the grounded analysis results. When participants rated the password level adequacy, the `$PasswordRating` was lowered by the Public Wi-Fi network level, even when the password level was strong. Similarly, the `$TimerRating` was lowered by the use of Public Wi-Fi or VPN over unencrypted Wi-Fi. When the `$NetworkType` changes to Public Wi-Fi, respondents rate the strong password and auto-logoff timer as *less than adequate,* because participants likely view these two requirements as reinforcements that raise the general level of security, but do not mitigate the threat. In our grounded analysis, we further saw participants focusing their attention on providing requirements to replace the weak network. One participant stated that the timer, password, and SSL are no longer effective, if the communication is happening over a vulnerable network like Public Wi-Fi. Another participant explained how, despite the use of employer's VPN, a public unencrypted Wi-Fi could still be vulnerable. In addition, our multi-level modeling results for the `$ConnectionRating` show that for the `Man-in-the-Middle` threat, participants generally rated SSL near adequate, but the ratings dropped in the presence of Public Wi-Fi. Moreover, we saw participants providing requirements refinements for SSL regardless of change in dimensions' levels. For example, five participants suggested to *update the SSL version*, and five participants suggested to *verify SSL certificates* and they replicated these modifications for all four-network types. Since the `$Connection` dimension in our vignette design is a blank dimension with one SSL level, we cannot derive conclusions on how raising the SSL security level would affect the other composite requirements in the scenario. However, we do plan on modifying our vignette design to account for these and other elicited mitigations to test their interaction with other levels.

The suggested refinements for SSL levels indicate that our proposed vignettes are incomplete, and that we should broaden the scope of our composition to include new dimensions/levels than what we proposed. Our grounded analysis also confirms that there are more dimensions to consider, such as *browser security configurations*. Secure communication relies on the browser's configuration, as we found 17 browser security reinforcements that 11 participants proposed as mitigations to increase the overall security level. Among these, seven browser security reinforcements were suggested in the presence of the employer's network and/or VPN over Encrypted Wi-Fi. After examining all the mitigations provided by these participants, we found that when `$NetworkType` is weak, participants focus their attention on replacing it because it significantly increases the security risk. When the risk is lowered by using a more secure `$NetworkType`, participants began looking at other dimensions to increase the overall security level.

In Section V, we discussed how experts identified ambiguous requirements proposed to reinforce, replace, and/or refine these requirements. The vignette dimensions were observed to affect participants' risk perception leading them to list mitigations based on the dimensions and their levels. For example, participants focus attention on replacing weaker requirements with stronger levels (e.g. replacing Public Wi-Fi), and that explains the high number of replacement mitigations provided for public Wi-Fi (see Table VIII). In addition, out of the total 907 mitigations, only 78 (9%) were not directly related to our dimensions in the study as they include categories such as browser security and device identifiers (see Fig. 2 for categories). Regarding ambiguity, in Section V we note that participants might assume that the public Wi-Fi is unencrypted, because vignette description omits mention of encryption. Similarly, the vignette does not provide details about the SSL dimension and participants made their own assumptions that made them list mitigations of refinements (e.g. version update), reinforcement, (e.g. certificate verification), and even replacement (e.g. TLS). This observation suggests two things with regards to ambiguity resolution: 1) when participants make assumptions to resolve ambiguity, they might lean towards assuming lower security (e.g. unencrypted Wi-Fi, insecure SSL versions); and 2) adding and removing requirements in a composition can have interactions by increasing or decreasing levels linked to the refined requirement. The method we introduce in this paper allowed us to assess such composition, however, additional work is needed to evaluate the effect of these elicted mitigations on the overall and dimension-specific risk perceptions.

VII. Related Work

We now review related work in requirement engineering relating to security and risk. Haley et al. introduced trust assumptions, which concern the extent to which security analysts trust domain properties [15]. Domain properties are those properties that engineers typically assume are true; thus, if they appear untrue the system will often fail. Because domain knowledge is often distributed across experts, there is a need to model consensus understanding of how and when trust assumptions can fail. Moreover, trust assumptions may interact; in which case, weaknesses in one assumption can impact the risk perception of another assumption.

Gandhi and Lee use multi-dimensional correlations to determine the criticality of a class of security constraints on the overall secure system behavior, thus examining the issue of interactions [10]. The method involves goal-driven scenario identification, keyword search and manual filtering and categorization of security requirements to curate an analysis pool and create a correlation model using formal concept analysis (FCA). The approach requires analysts to express the correlations as logical formula *a priori* to realizing how requirements compose. One limitation of this approach is effort required to curate the analysis pool and the lack of consensus across experts to inform the claims of compositionality.

A complementary approach is Franqueira et al.'s risk-based argumentation framework based on Haley et al. [15] that guides analysts to construct arguments from grounds, warrants, claims and rebuttals [8]. The approach yields prioritized risks that compose mitigations under the argument structure based on the analysts' reasoned prediction of risks and their mitigations. In contrast, Zarghami et al. describe a risk-identification process that helps analysts explicate assumptions underlying service providers to better understand the risk of their composite application [27]. When these assumptions fail,

the composition can break due to loss or degradation of service. This approach supplements the approaches by Haley et al. and Franqueira et al., because it treats services as opaque boundaries.

Cailliau and van Lamsweerde introduce a framework for goal-oriented risk analysis that assigns probability estimates to goal satisfaction in view of obstructions [5]. They compute obstacle combinations based on independent severity levels to select the most appropriate mitigations; however, this approach does not account for analyst perceptions of risk, or for how requirements compose to reduce risk.

Hibshi et al. studied security analysts to understand how they make decisions based on Endsley's Situation Awareness model [13, 14]. In this work, the role of context in directing the analyst's attention and mitigations surfaced as critical challenge: the variability of perceivable contexts and threats led analysts to draw dramatically different conclusions. In this paper, we fix the threat and context to target more precisely the expert's ability to perceive risk and attribute security adequacy to components and their interactions.

VIII. CONCLUSIONS

In this paper we introduce a method to study how changes in security requirements composition affects experts' risk perception. We show results of running 64 factorial vignettes on 174 participants and two threat scenarios. With our approach, we were able to use security expert's judgment to evaluate joint requirements interactions in a security scenario by measuring the security adequacy ratings of the overall scenario and for the individual requirements composed in the scenario. Our method also allows for extracting weights for the individual requirements in order to understand their assigned priorities in the scenario. In addition, we show how our method is effective in eliciting three categories of security requirements modification: refinements, replacements and reinforcements. In the future, we plan to adapt our method to analyze scenarios that have a more complex attack surface to better understand how composition changes.

ACKNOWLEDGMENT

We thank our study participants and Dr. Jennifer Cowley at SEI-CERT who consulted on the user study design. This research was funded by National Security Agency, and the Software Engineering Institute.

REFERENCES

[1] C. S. Alexander and H. J. Becker, "The use of vignettes in survey research," *Public Opin. Q.*, vol. 42, no. 1, pp. 93–104, 1978.

[2] K. Auspurg and T. Hinz, *Factorial Survey Experiments*, vol. 175. SAGE Publications, 2014.

[3] D. Bates, M. Maechler, B. Bolker, S. Walker, R. H. B. Christensen, H. Singmann, and B. Dai, *lme4: Linear mixed-effects models using Eigen and S4*. 2014

[4] J. Cohen, *Statistical Power Analysis for the Behavioral Sciences*. L. Erlbaum Associates, 1988.

[5] A. Cailliau, A. van Lamsweerde. "A Probabilistic Framework for Goal-Oriented Risk Analysis," 22nd IEEE Int'l Req'ts Engr. Conf., pp. 201-210, 2014.

[6] Andy Ellis, "SSL is dead, long live TLS - The Akamai Blog," 14-Oct-2014. [Online]. Available: https://blogs.akamai.com/2014/10/ssl-is-dead-long-live-tls.html. [Accessed: Mar-2015].

[7] F. Faul, E. Erdfelder, A.-G. Lang, and A. Buchner, "G* Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences," *Behav. Res. Methods*, vol. 39, no. 2, pp. 175–191, 2007.

[8] V.N.L. Franqueira_, T.T. Tuny, Y. Yuy, R. Wieringa, B. Nuseibeh. "Risk and Argument: A Risk-Based Argumentation Method for Practical Security," 19th IEEE Int'l Req'ts Engr. Conf., pp. 239-248, 2011.

[9] A. Gelman and J. Hill, *Data analysis using regression and multilevel/hierarchical models*. Cambridge Univ. Press, 2006.

[10] R.A. Gandhi, S-W. Lee. "Discovering and Understanding Multi-dimensional Correlations among Certification Requirements with application to Risk Assessment," 15th IEEE Int'l Req'ts Engr. Conf., pp. 231-240, 2007.

[11] B. Glaser, Theoretical Sensitivity: Advances in the Methodology of Grounded Theory. Sociology Press, 1978.

[12] T. McGuiness, "Defense in depth," *SANS*, 2001.

[13] H. Hibshi, T. Breaux, M. Riaz, and L. Williams, "Towards a framework to measure security expertise in requirements analysis," in *2014 IEEE 1st Workshop on Evolving Security and Privacy Req'ts Engr. (ESPRE)*, pp. 13–18, 2014.

[14] H. Hibshi, T. Breaux, M. Riaz, and L. Williams, "Discovering Decision-Making Patterns for Security Novices and Experts," Inst. for Softw. Research., Carnegie Mellon Univ., Pittsburgh, PA, Tech. Rep. CMU-ISR-15-101, Mar 2015.

[15] C. B. Haley, R. C. Laney, J. D. Moffett, and B. Nuseibeh, "The effect of trust assumptions on the elaboration of security requirements," 12th IEEE Int'l Req'ts Engr. Conf., pp. 102–111, 2004.

[16] Marshall Honorof, "SSL vs. TLS: The Future of Data Encryption," 06-Sep-2013. [Online]. Available: http://www.tomsguide.com/us/ssl-vs-tls,news-17508.html. [Accessed: 08-Mar-2015].

[17] M. Howard, J. Pincus, and J. M. Wing, *Measuring relative attack surfaces*. Springer, 2005.

[18] G. Jasso, "Factorial survey methods for studying beliefs and judgments," Sociol. Methods Res., vol. 34, no. 3, pp. 334–423, 2006.

[19] "NIST/ITL Special Publication (800)," 02-Jan-2015. [Online]. Available: http://www.itl.nist.gov/lab/specpubs/sp800.htm. [Accessed: 02-Jan-2015]

[20] OWASP, "OWASP Top Ten Project - OWASP," 30-Dec-2014. [Online]. Available: https://www.owasp.org/index.php/Category:OWASP_Top_Ten_Project. [Accessed: 30-Dec-2014].

[21] R Core Team, *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing, 2013.

[22] P. H. Rossi and S. L. Nock, *Measuring Social Judgments: The Factorial Survey Approach*. SAGE Publications, 1982.

[23] J. Saldaña, *The coding manual for qualitative researchers*. Sage, 2012.

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

[25] US-CERT, "OpenSSL 'Heartbleed' vulnerability (CVE-2014-0160) | US-CERT," US-CERT, 08-Apr-2014. [Online]. Available: https://www.us-cert.gov/ncas/alerts/TA14-098A. [Accessed: Mar-2015].

[26] L. Wallander, "25 years of factorial surveys in sociology: A review," *Soc. Sci. Res.*, vol. 38, no. 3, pp. 505–520, Sep. 2009.

[27] A. Zarghami, E. Vriezekolk, M.Z. Eslami, M. van Sinderen, R. Wieringa. "Assumption-Based Risk Identification Method (ARM) in Dynamic Service Provisioning," 21st IEEE Int'l Req'ts Engr. Conf., pp. 175-184, 2013.