Improving Security Requirements Adequacy

An Interval Type 2 Fuzzy Logic Security Assessment System

Hanan Hibshi[1,2], Travis D. Breaux[1], and Christian Wagner[3,4]
Institute for Software Research, Carnegie Mellon University, Pittsburgh, Pennsylvania, USA[1]
College of Computing, King Abdul-Aziz University, Jeddah, Saudi Arabia[2]
Institute of Computing & Cyber Systems, Michigan Technological University, Houghton, Michigan, USA[3]
Lab for Uncertainty in Data and Decision Making, School of Computer Science, University of Nottingham, Nottingham, UK[4]
{hhibshi, breaux}@cs.cmu.edu, chwagner@mtu.edu

***Abstract*— Organizations rely on security experts to improve the security of their systems. These professionals use background knowledge and experience to align known threats and vulnerabilities before selecting mitigation options. The substantial depth of expertise in any one area (e.g., databases, networks, operating systems) precludes the possibility that an expert would have complete knowledge about all threats and vulnerabilities. To begin addressing this problem of fragmented knowledge, we investigate the challenge of developing a security requirements rule base that mimics multi-human expert reasoning to enable new decision-support systems. In this paper, we show how to collect relevant information from cyber security experts to enable the generation of: (1) interval type-2 fuzzy sets that capture intra- and inter-expert uncertainty around vulnerability levels; and (2) fuzzy logic rules driving the decision-making process within the requirements analysis. The proposed method relies on comparative ratings of security requirements in the context of concrete vignettes, providing a novel, interdisciplinary approach to knowledge generation for fuzzy logic systems. The paper presents an initial evaluation of the proposed approach through 52 scenarios with 13 experts to compare their assessments to those of the fuzzy logic decision support system. The results show that the system provides reliable assessments to the security analysts, in particular, generating more conservative assessments in 19% of the test scenarios compared to the experts' ratings.**

***Index Terms* —user study; vignettes; scenarios; recommender system; security requirements; fuzzy logic; type-2; uncertainty.**

I. Introduction

The rate of security attacks on different organizations has been increasing over recent years. According to the *Global State of Information Security* survey, security incidents increased in 2015 by 38% above reports from 2014, which led to a 56% increase in intellectual property theft [22]. The survey also reports that 53% of organizations are conducting employee training and awareness programs, and 54% designate a Chief Security Officer (CSO) to lead teams of security specialists [22]. The focus on establishing professional personnel to address this problem illustrates the reliance on human experts to comprehensively assess the security of systems. However, despite the abundance of security requirements, checklists, guidelines and best practices, such as the U.S. NIST Special Publication 800 Series, human analysts still face substantial challenges in the selection of the appropriate security requirements to mitigate threats. For example, depending on the chosen attack scenario, analysts must still evaluate a range of possible security authentication requirements, such as password complexity, single and multifactor authentication.

In addition to the composition of requirements, the number of security experts in the world is scarce. According to the U.S. Bureau of Labor statistics, there are 82,900 information security analysts in the U.S. in 2014, earning a median of $89,000 a year [29]. In 2014, Cisco's Annual Security report warned that the shortage in security professionals by the end of 2014 is reaching one million [4]. In addition, more than 209,000 cybersecurity jobs in the U.S are unfilled and 53% growth in demand is expected by 2018 [27]. The scarcity of experts and the need for cyber security, makes the provision of intelligent decision support and semi-automated solutions a necessity.

The contribution of this paper is a novel empirical method for constructing an Interval Type-2 Fuzzy Logic System (IT2FLS) for automated cyber security assessments. The method extends a scalable technique for acquiring adequacy ratings of security requirements by measuring the extent to which these requirements interact to affect security, while accounting for uncertainty across raters [12]. The use of fuzzy sets (FSs) associated with linguistic labels, in combination with fuzzy logic rules provides the resulting decision support system with a high degree of human (expert) interpretability, which in turn is vital for its evaluation and acceptance. The paper specifically adopts Type-2 FSs which offer an advantage over Type-1 FSs, because they allow the distinct capture of both inter- and intra-expert uncertainty [31]. In this paper, we present a series of studies to construct and evaluate the IT2FLS in using a series of scenarios.

The remaining paper is organized as follows: in Section II we show background on the process of security assessment, general aspects of uncertainty and IT2FLSs; in Section III, we present our overall approach to the problem; in Section IV, we present the rule base extracted from user surveys; in Section V we present evaluation results; in Section VI we discuss threats to validity; we discuss and conclude in Section VII.

II. Background

We now provide background to our interdisciplinary approach to model security assessments.

A. Security Assessment

A security assessment is a decision by a security analyst about a system's readiness to withstand potential cyber-attacks. The ISO 27000 Series, U.S. NIST Special Publication (SP) 800 Series, and the Information Technology Infrastructure Library (ITIL) requires analysts to conduct risk assessments to determine readiness. Under NIST SP 800-53, the analyst decides if a specific system is high, medium or low impact and then the analyst satisfies the impact rating by selecting security controls (e.g., audit events, lock sessions, etc.) Each control represents a class of technology aimed at mitigating a security threat. An unaddressed challenge to this approach is that the threats, vulnerabilities and security mitigations change over time: new technologies introduce new vulnerabilities, and attackers become more sophisticated in their attacks to weaken a collection of mitigations that work together. Security auditing tools, such as the Security Content Automation Protocol (SCAP), trace vulnerabilities to mitigations; however, these tools rely heavily on vendors to supply vulnerability data, and they do not address design risks from custom built software.

We highlight general challenges to human-based security assessments:

- *Context*: experts' risk assessment of a system must consider the system context in which the requirements apply [12, 13].
- *Priorities*: some requirements have higher priorities than others, depending on their strength in mitigating threats [12].
- *Uncertainty*: security risk assessment and decision-making includes a level of uncertainty [13, 23].
- *Stove-piping*: security expertise crosses different domains of knowledge such as hardware, software, cryptography, and operating systems [13].

Our aim is not to remove the above challenges through increased decision-support. Instead, we account for these challenges by modeling human decision making with uncertainty, in a security assessment support tool based on collected data from various security experts.

B. Uncertainty in Requirements Engineering

Uncertainty is increasingly a focal point for researchers in requirements and software engineering. In architecture, Garlan argues that the human-in-the-loop, mobility, rapid evolution, and cyber physical systems are possible sources of uncertainty [9]. Esfahani and Malek identify sources of uncertainty in self-adaptive systems and they include the human-in-the-loop as a source of uncertainty [6]. In requirements engineering, Yang et. al [35] used machine learning to capture language uncertainties in speculative requirements. The approach succeeds at identifying speculative sentences, but performs weaker at identifying the scope of uncertainty when identifying specific parts of speech such as adjectives, adverbs, and nouns. The FLAGS is a goal modelling language introduced to model uncertainty in self adaptive systems [1, 21].

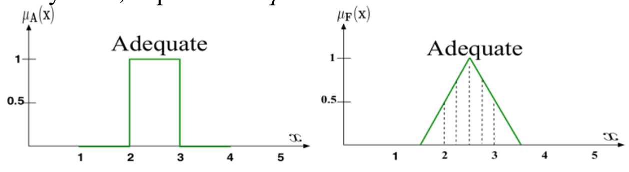
Cailliau and van Lamsweerde introduce a method to encode knowledge uncertainties in probabilistic goals [2]. This method characterizes uncertainty as the probability of goal satisfaction using estimates of likelihood collected from experts. Although the authors' method is sound, the reliability depends heavily on a third-party method to record expert estimates [2]. In this paper, we contribute a novel method to elicit estimates and incorporate estimates into an IT2FLS.

C. Uncertainty and Type 2 Fuzzy Logic

Zadeh introduced Fuzzy Logic (FL) in 1965 as a mathematical tool in which the calculations use a degree of truth rather than simple propositions; true or false [36]. To illustrate, security experts have been shown to use the linguistic adjectives *inadequate*, *adequate*, and *excessive* on a 5-point semantic scale to evaluate the security of the scenarios [12]. Let X be our universe of discourse on a continuous, real-valued, inclusive scale $X = [1,5]$ and set $A \in X$ to represent "Adequate." Assume that an interval between $[2,3]$is adequate, as shown in Fig. 1(a). The function $\mu_A(x)$ is the membership function (MF) to describe *A*, where 1 is true and 0 is false:

$$A \Rightarrow \mu_A(x) = \begin{cases} 1 & 2 \le x \le 3 \\ 0 & otherwise \end{cases} \qquad (1)$$

Based on the definition above, a value like 1.9 for example is not *adequate*, because 2 is the inclusive threshold value for the *adequate* set, but 3.1 is very close to *adequate* or is *adequate* with a lesser degree than 1, but greater than 0. To address this concern, fuzzy set theory allows one to express *to what degree* a value *x* belongs to a fuzzy set [16, 36]. Figure 1(b) shows how a fuzzy set *F,* captures *adequate*.



(a) Crisp sets

(b) Fuzzy sets

Fig. 1. The definition of adequate in crisp and fuzzy sets

A fuzzy set *F* of values in *X* may be represented as a set of ordered pairs of the value *x* and its membership grade [16],

$$F = \{(x, \mu_F(x)) | \, x \in X\} \qquad (2)$$

Type-1 MFs summarize the results of experts' ratings into a single MF, suppressing the uncertainty in the data. Alternatively, Type-2 MFs model the uncertainty by providing a *footprint of uncertainty* (FOU) [16, 19]. Figure 2 below highlights a prototypical Type-2 FS, specifically, a so-called Interval Type-2 (IT2) FS, which is completely defined by an upper and lower Type-1 MF which together form the FOU. Note that IT2 FSs are a simplification of General Type-2 FSs, where the former assign the value of 1 to all secondary memberships (i.e. all points on the FOU are weighted equally as 1) and the latter allow variation in[0,1]of this weighting [16,17]. This paper only uses IT2 FSs throughout.

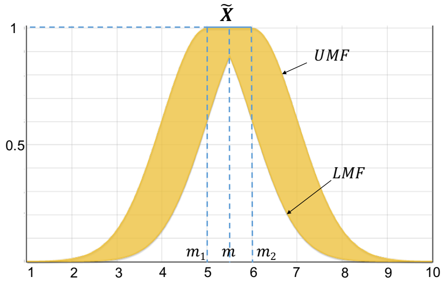


Fig. 2. Type-2 FOU constructed by blurring a Type-1 MF

D. Interval Type 2 Fuzzy Logic Systems

Type-2 FSs are used in rule-based intelligent systems. The rule base is expressed as a collection of if-then statements and they can be collected by surveying experts in the field [17]. In the remainder of this paper we will show how we build a security system using an IT2FL approach. Figure 3 shows the main components of the proposed system. The components shown in Fig. 3 represent what is typically found in IT2FLS [17, 19]. The components in an IT2FLS are similar to a Type-1 FLS, but with the addition of a type reducer. The type reducer reduces the inference engine's IT2FS output to an interval Type-1 FS that the defuzzifier can use to produce the final crisp output number.

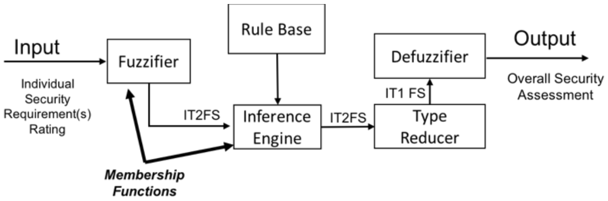


Fig. 3. Mamdani IT2FLS for Security Assessment

III. Overall Approach

In this section we explain our overall research method to build a security assessment system using IT2FLS. Our contribution is two-fold:

- A comprehensive approach for developing the linguistic labels and associated membership functions for an FLS.
- An innovative approach to designing the rule base from surveys of domain experts.

Now, we describe these two contributions.

A. Developing Linguistic Labels and Associated Fuzzy Sets

For the FSs used in the security assessment system, we had to decide on the appropriate linguistic labels, which are the vocabulary used in the system. The choice of labels relies on background knowledge and expertise in the field, and user surveys that support the choices made [16, 19]. We first conducted a focus group using 5 researchers in our lab. A group of labels to describe overall security levels were discussed in the context of a concrete scenario. The labels discussed in the focus group include labels used in prior research [16, 19] in fuzzy logic. An outcome of this focus group was the recognitions that in security domain, requirements serve to mitigate threats and decrease risk. With the goal of mitigating threats, security requirements can be described as: inadequate, adequate, or excessive, because security requirements are often *cost* requirements, meaning the value is not so obvious to achieve primary system goals and users often have difficulties seeing the benefits. Furthermore, excessive security has negative financial and usability effects, while adequate security is what an organization might settle for. Hence, we developed three labels to describe security adequacy: inadequate, adequate, and excessive.

Next, we will describe how we evaluated these labels experimentally, and the English language proficiency test we conducted on each participant, so we can reliably use the labels' in our design and evaluation.

1) Evaluating the Linguistic Labels Experimentally

It is important to realize here that we are creating a *new scale* to measure our *construct of security adequacy*, because there are no existing, empirically valid scales to measure this construct. Psychometric researchers describe this type of scale as an *ad hoc scale* due to the lack of valid or reliable scales [8]. Creating ad hoc scales requires evaluation to examine the reliability of the scales rather than relying on the face validity alone [8]. In contrast to construct validity, the face validity is subjective: if a test or measure *looks like* it will measure what it is supposed to measure, then it is said to have face validity [18].

We bootstrapped our scale terminology by selecting standard dictionary synonyms to the three main anchor points, inadequate, adequate, and excessive, to yield a 17-word dataset. We replaced *adequate* with *average*, because we were interested to see where *excessive* and *inadequate* rank compared to the average or mid-point. We surveyed 205 participants on Amazon Mechanical Turk (AMT) asking them to rank the 17 words. Our results show that *excessive* ranks higher than "average" across four scenarios, and *inadequate* ranks below average in three out of four scenarios [11].

English proficiency is necessary to yield accurate results as a participant need to have the language capabilities to distinguish meanings among the list of words that contain synonyms. One could limit the AMT participation to people located in the U.S., however, location is unreliable, because U.S. residency does not guarantee English proficiency, and there are ways to fake locations. We tested English language reading proficiency using a subset of the Nelson-Denny reading comprehension test. We discuss these results in detail, along with the word ranking study in a separate technical report [11]. In the word-ranking study, participants are only allowed to proceed to the ranking survey, if they pass the proficiency test in 15 minutes with a score of 80% or above.

2) Eliciting the Membership Functions for the Fuzzy Sets

The membership function definition depends upon a scale assignment along an interval (e.g., from one to ten) for each word selected from our ranking study. We adopted the approach commonly accepted by the fuzzy logic research community in which experts are asked to assign the interval start and end points on one scale for each word [16, 19]. Participants were asked to specify the intervals of the 17 words from our previous ranking survey plus the word adequate (total 18 words) using the text template we show below, replacing *Adequate* with each of the other 17 words. We include a security scenario to add context to each word as follows:

> A security expert was asked to rate a security scenario with regards to mitigating the **Man-in-the-Middle** threat.
>
> The expert would give an overall security rating using a linguistic term.
>
> In the next sections of this survey, we will present 18 linguistic terms describing the overall security of a scenario. We would like you to mark an interval between 1-10 that represents each term.
>
> **Note:** Intervals for different terms can overlap.

For each word (e.g., "adequate"), participants were asked:

> Imagine **"Adequate"** represented by an interval on a range from 1-10. Where would you indicate the start and end of an **"Adequate"** security rating?

We randomized the word order in the survey, and we recruited participants by sending out email invitation to security mailing lists at Carnegie Mellon University. Similar to our prior work [12], we use a test to assess the security knowledge of the survey participants.

We collected intervals from 38 security experts that consists of 74% males, 18% females, and 8% unreported. The average score on the security knowledge test is 6 out of 10 possible points ($sd = 1.75$). For each word, we calculated the average for the interval end points that we collected from participants. The results show that the three words: inadequate, adequate, and excessive are sufficient to be used as FSs covering an interval from 1-10. Figure 4 shows the selected FSs and their coverage of the 1-10 interval. The solid region represents the interval between the mean values of the start and end points collected from the experts. The shaded region on each side of the solid region represents the standard deviation for that point, which represents the uncertainty surrounding the mean value. It is only possible to cover the entire region from 1-10 because of the uncertainty that yield overlapping intervals for the three words. Mendel explains how this approach improves performance as it reduces the size of the rule base [16].

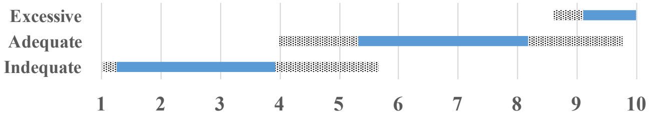


Fig. 4. The fuzzy sets with the start and end means and standard deviation

After choosing the labels for the fuzzy sets, we now explain how to derive the MFs. We create the Type-1 MF and then *blur* its mean by adding a degree of uncertainty and creating the shaded region that represents the FOU. We calculate the mean for the Gaussian Type-1 MF by averaging the two end points for the interval representing each word: $Mean_{interval} = (Mean_{start} + Mean_{end})/2$. Then, we average the standard deviation : $\sigma_{interval} = \sqrt{(\sigma_{start}^2 + \sigma_{end}^2)/2}$.

To represent the uncertainty level surrounding the Gaussian Type-1 MF: let α represent the uncertainty level, and then calculate two means: $m1$ and $m2$ and use these for the upper and lower membership calculations: $m1 = Mean_{interval} - \alpha$, and $m2 = Mean_{interval} + \alpha$. We assume that we have 50% uncertainty present in our data, which makes: $\alpha = 0.5$. Table I shows the final means and standard deviations for each word label for fuzzy sets. Figure 5 illustrates the membership functions. We use the same MFs for all the inputs: network, SSL, password, and timer; and for the output.

B. *Discovering the Fuzzy Rules from the Experts*

We need a rule base to build our security assessment system. The rule base should closely emulate how a human expert makes decisions. While security experts understand the domain and can make relevant decisions [13], it is unclear the extent to which this knowledge exists in the form of rules, or how easily experts recall relevant knowledge to express rules. The canonical approach is to develop a set of if-then rules and then ask experts to evaluate the rules and their consequences. This approach has been followed by fuzzy logic researchers [16, 19].

We choose a different approach. We believe asking experts directly about rules puts security requirements into a checklist, which treats the requirements as independent. In prior work [12], we found that requirements exist in composition with priorities among them. In addition, if we select a rule base based on our own judgment and then present the rules to experts to evaluate, we would introduce our own knowledge bias. How can we know that we selected the rule set that is representative of the problem, and how many antecedents exist? Wu and Mendel suggest that the number of rules could increase exponentially to the number of membership functions for each input [34]. We discuss this further when we present our extracted rules in Section IV.

TABLE I. SUMMARY STATISTICS OF THE THREE FUZZY SETS

Word	$\boldsymbol{Mean_{Interval}}$	$\boldsymbol{\sigma_{Interval}}$	$\boldsymbol{m_1}$	$\boldsymbol{m_2}$
Inadequate	2.58	1.26	2.08	3.08
Adequate	6.75	1.75	6.25	7.25
Excessive	9.50	1.35	10.00	9.00

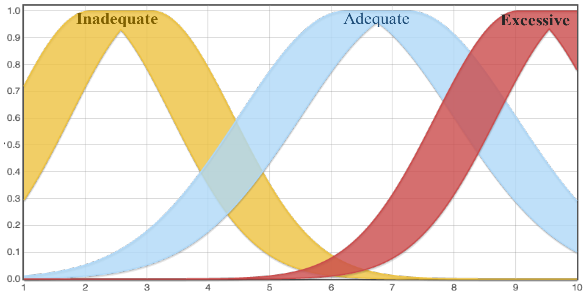


Fig. 5. The MFs for the input/output variable(s)

C. *Designing the Fuzzy Logic Assessment system*

A number of researchers have built software packages and tools for IT2FLSs [3, 20, 33, 34]. Packages and tools were designed for the mathematics modeling and simulation software MATLAB, and are based on the .m files originally written by Mendel and Wu. We chose to use the *Juzzy* and *JuzzyOnline* Java-based toolkit to obtain our results, because these are open-source and actively maintained by a team of fuzzy logic researchers [30, 32]. Based on prior IT2FLSs research [34], we made the following design choices:

1) Input and output MF shapes: The choice of MFs is dependent upon the context of the problem and other factors, such as continuity, and computational cost. We chose to use a Gaussian shape for our MFs for it's added advantage of simplicity and faster computation time [34]. As explained in Section III.A.2, we selected three membership functions for each input domain: inadequate, adequate, and excessive.

2) Input Fuzzification: An important step in a fuzzy system is to fuzzify the input by mapping an input vector $X = (x_1^{'}, \ldots, x_p^{'})$ into p fuzzy sets $X_i , i = 1,2, \ldots, p$ [16, 34]. We choose to use the singleton fuzzifier, where: $\mu_{X_i}(x_i) = 1$ at $x_i = x_i^{'}$ and : $\mu_{X_i}(x_i) = 0$ otherwise. Singleton fuzzifiers are more practical due to their simplicity [16, 34]. The input to the system would be a number between 1-10 representing the level of the security requirements adequacy to mitigate a threat.

3) Rules: we construct the rules following the Mamdani 'style', thus preserving the use of linguistic labels at the output stage – supporting human-interpretability [16, 34]. We also chose the minimum t-norm, because we want our security assessment system to act conservatively.

IV. RULESET DISCOVERED FROM SECURITY EXPERTS

In this section we explain how we translate the user survey results from prior work [12] to a rule base for the assessment system. In the user study, participants rated the overall security of scenarios using a 3-point scale: 1=excessive, 0=adequate, and -1=inadequate. Participants also rated four individual requirements related factors in scenarios: network type, using SSL, password strength, and presence of a timer using a five-point semantic scale with: 5=excessive, 3=adequate, and 1=inadequate with the midpoints: 2,4 between inadequate-adequate and adequate-excessive, respectively. Experts rated four security scenarios with four network types: employer's network, public Wi-Fi, unencrypted VPN, and encrypted VPN. Each scenario included a password, timer, and SSL requirements. The password and timer had two conditions each (either strong or weak) [16].

Based on our prior results [16], we built the rules as follows:

- The regression results for the overall security ratings indicate network type has the major significant effect, it takes priority over other requirements.
- The network rating suggests that network type can drop the overall ratings significantly with no significant effect for the other factors, hence it is safe to remove the other factors from the rule antecedents only when the network type drops to inadequate.
- When network type increases to adequate, other requirements are included as antecedents, because the statistical results show that the model with all the four factors exhibits an effect over the null model.

Next, we show how we applied the above heuristics.

A. The Inadequate Network

The *public Wi-Fi* and the *VPN over unencrypted Wi-Fi* networks significantly dropped the overall security ratings towards inadequate (Public Wi-Fi: $Mean = -0.7$, VPN-unencrypted $Mean = -0.4$). The public Wi-Fi ratings are closer to inadequate ($Mean = 1.3$) while VPN-unencrypted ratings are in between adequate and inadequate ($Mean = 2.2$) [16]. From the above, we can infer that when the network type is inadequate, then no other requirement(s) would matter in deciding the adequacy of the overall security. Hence, we construct the following rule:

$$R^1: IF\ \textbf{NetworkType}\ is\ \textbf{Inadequate}$$
$$THEN\ \textbf{OverallRating}\ is\ \textbf{Inadequate}$$

Reduction of rules in rulesets used in intelligent systems simplifies the reasoning for the human analysts interacting with the system [21]. Without the results from our user study [16], we would have more rules with more input combinations. For example, we would have a four antecedent rule, wherein each input antecedent has three MFs: inadequate, adequate, and excessive. For the inadequate network alone, input combinations of the remaining three inputs (SSL, password and timer) will result in 27 rules and, if we follow a canonical approach, we would need to survey experts to obtain the consequents of all 27 rules. However, our approach derives rules from the statistical analysis of the empirical results in which factor levels that are not significant are dropped.

B. The Adequate and Excessive Network Types

When the network adequacy level increases, then the rules for factors would change as well. The remaining network types in the study: *employer's network* and *Encrypted VPN* were rated close to adequate, but never close to excessive ($Mean = 2.6$, and $Mean = 2.9$ respectively). The overall security of the scenario was rated below adequate: ($Mean = -0.19$, and $Mean = -0.16$ respectively). This data is not sufficient to infer a rule similar to R1; i.e. we cannot use network adequacy alone in a single antecedent rule. However, our data does show that when the network adequacy level improves, participants begin paying attention to the other factors in the scenario and their decisions become based on the composition of these other factors. The regression model for the overall security rating shows that all the factors in the scenario are predictors of the model [12]. Hence, we decide to include more input variables in our rule set antecedents. Table II below shows the antecedents and consequent combinations for the remaining rules that we constructed from our scenarios. The column *R#* is the rule number, *Antecedents* are the requirements that serve as input antecedents in the if-then rules, *Con.* is the consequence output that is the rating of the overall security.

From Table II, we can see how priority is given to the network, once it is adequate then other factors are considered. Only when all other factor levels are adequate, does the overall security rise to adequate. Our results show that certain factors have weaker ratings, and that strengthening the requirements without strengthening the network does not improve the overall security rating.

Since our results did not show significant effects for excessive ratings, it becomes harder to infer rules for excessive cases. Regardless, we include the extreme case where all the security requirements are rated as excessive in order to raise the overall security to excessive. Combinations that are absent from the rules in the table, such as adequate/ excessive combinations, did not have any statistical significant results in the dataset.

TABLE II. RULES FOR SECURITY ASSESSMENT SYSTEM

R	**Antecedents (IF)**				**Con. (THEN)**
#	***Network***	***SSL***	***Password***	***Timer***	***Overall***
R1	I				I
R2	A	I			I
R3	A		I		I
R4	A			I	I
R5	A	A	A	A	A
R6	E	E	E	E	E

V. QUALITATIVE SYSTEM EVALUATION

In this section we explain our qualitative approach for system evaluation. Qualitative methods are better suited to our system evaluation as we are looking for the participants' description of the process, the rationale, and their reasoning.

A. Evaluation Process

We evaluate our IT2FLS using a two stage process: first, we survey 13 experts and have each evaluate 4 scenarios each (52 total test scenarios); and second, we conduct follow-up interviews to discuss participants' decisions and their rationale.

The survey used in the first stage is similar to the survey in the original factorial vignettes study, which examined the interaction of a public/private network with SSL-encrypted connections, varying password strengths and an automatic, timed logout feature [12]. We modified the original design to limit the network levels to two levels that highly contrast each other: public unencrypted Wi-Fi and encrypted VPN. Furthermore, we use two levels for the password (weak/strong), and two levels for the logout timer (no timer, 15 min timer). We did not use a number of different SSL levels as this would present an obvious focal point for the participant to become concerned about security [12]. Hence, we reworded the scenario to describe SSL as follows: "The browser is already using the latest (and patched) version of SSL/TLS for the session."

Participants were randomly assigned to different conditions, and we randomized the order in which they see different vignettes. Each participant rates four vignettes in total with combinations that show the two levels of each variable: network, timer, and password.

For each of the four scenarios, participants provide their overall security judgement of the scenario. Participants choose either: inadequate, adequate and excessive to evaluate the overall security adequacy without the use of any scales or numbers.

The four inputs to the IT2FLS are the adequacy ratings for the network, the SSL, the password, and the timer. We use the participants' provided ratings as inputs to our system. The output would be the overall security rating represented by a number on an interval from 1-10. After we calculate the output, we interview participants and remind them of their initial ratings including the overall security judgement of the scenario. Before showing them the output of the system, we ask them to describe the overall security ratings on a scale from 1-10, and why they rated a scenario the way they did. Then, we show the participant the output of our security assessment system in the form of fuzzy sets and we solicit their opinion. Finally, we ask participants to state what they would change in the scenario to improve the adequacy ratings, and in contrast, what would they imagine to be the worst possible change to drop the adequacy ratings further.

B. *Evaluation Results*

The participants median score on the knowledge test was 7 out of 10. Three out of 13 participants work in cybersecurity at federally-funded research and development centers, one participant has 10 years of experience as a security consultant, and the remaining 9 are graduate students from Carnegie Mellon University who completed security courses and who are involved in security research.

Table III shows the participant agreement for all eight scenario combinations: the network type, the password (Pwd), the logout timer (Timer), the total number of participants per scenario, and the percent agreement, which is the total number of overall ratings that match the ratings produced by the security assessment system. In Table III, we see participants disagreed with the system's overall security rating predictions. We conducted follow-up interviews with nine participants. Six participants agreed with the security assessment system's overall ratings for 4/4 scenarios: however, they explained that their assessment was *borderline* between two rating levels. Two participants agreed with 3/4 scenarios in the system's result: in the one disagreeable scenario, both participants provided an excessive rating while the system rated the scenario as adequate. Finally, the last participant P5, who scored 7 on the security knowledge test, disagreed with the system for 4/4 scenarios, because they mistakenly believed that SSL was an adequate mitigation against man-in-the-middle attack in all scenarios, even when the network is public Wi-Fi. The participant explains: "for the purpose of man-in-the-middle, SSL is all [that] we need; if we worry about sniffing while in a public place, then passwords and timers are important." The participant acknowledged why the overall security could be inadequate: "If we are worried that users may not understand insecure certificates, then the VPN over an encrypted connection might provide an extra layer of security."

TABLE III. PARTICIPANT AGREEMENT WITH OVERALL SECURITY

Scenario			**Total Participants**	**Agreement Ratio**
Network (Wi-Fi)	***Pwd***	***Timer***		
Public unencrypted	Weak	None	5	4/5 (80%)
Public unencrypted	Weak	15-min	8	6/8 (75%)
Public unencrypted	Strong	None	8	6/8 (75%)
Public unencrypted	Strong	15-min	5	3/5 (60%)
VPN over encrypted	Weak	None	8	6/8 (75%)
VPN over encrypted	Weak	15-min	5	2/5 (40%)
VPN over encrypted	Strong	None	5	2/5 (40%)
VPN over encrypted	Strong	15-min	8	4/8 (50%)

The follow up interviews helped us verify the participants' inputs, check for mistakes, and identify false positives. By false positives, we mean that participants could provide assessments that match the results of the system, but their reasons and priorities for security requirements did not match what the rule base had encoded. We found one false positive, participant P8, who scored 8 on the security knowledge test. Unlike P5 who disagreed with the system, P8 agreed but using a rationale similar to P5. Participant P8 mistakenly believed that SSL made the other factors less relevant, because they believe that SSL alone is sufficient to defeat man-in-the-middle attacks. The participant did not rate SSL as adequate, because they were concerned about checking the certificates and about whether or not users would trust untrusted certificates.

We asked participants: "what is the most important change in the scenario that, if it occurs, will cause you to drop your ratings?" All eight participants identified SSL, which only had one level; participants did not see stronger or weaker SSL variants, despite the existence of such variants. Participants identified requirements when weaker settings were presented: e.g., if they saw no timer, they would suggest adding a timer. This behavior was expected, because participants saw combinations where they reviewed both weak and strong settings for network, timer, and password.

We asked participants to identify requirements changes that would cause them to improve their adequacy ratings. Participant P8 indicated they would improve SSL by ensuring the server certificates are checked. The remaining six participants all suggested avoiding public unencrypted Wi-Fi and replacing it with a VPN over encrypted Wi-Fi or even better, as two participants suggested, using their own private home network. The six participants also suggested using a timer for automatic logout instead of no-timer, and using a stronger password setting instead of a plain 8-character password with no enforced character requirements. One participant suggested adding two-factor authentication to the scenario.

Our survey results reveal when participants provide different ratings to the same requirement level in two different scenarios. Eight of 13 participants provided different network ratings for the same network, but in two different scenarios. Four of eight participants clarified their choice during a follow-up interview. Three of four participants reported not remembering their previous choice, which suggests within-subject's variance. The remaining participant reported providing different network ratings, because they believe that their decisions were impacted by other requirements settings, such as the timer and password.

VI. Threats to Validity

When reporting results from surveys and experiments, it is important to address threats to validity that arise from the study design, study execution and interpretation of results [26]. The threats to validity for the previously collected vignette study data are reported, previously [12]. We now review the threats for the new studies reported in this paper.

Construct validity concerns how well the measurements we take correspond to the construct of interest [26]. As shown in Section III, we conducted a series of studies to evaluate our ad hoc scales, including the word rank study and word interval study to determine the adequacy label intervals. To ensure that participants have a shared understanding of the ratings, we provided one-sentence definitions for each rating level. For the experience indicator variable ($Score), we are the first to introduce such a test, thus we need further evaluation to assess the level of expertise indicated by this score.

Internal validity is the degree to which a causal relationship can be inferred between the independent predictor variables and the outcome dependent variables [26]. In the word interval study to collect the start and end points for the word labels, we randomize the order of words shown to each participant. In the evaluation survey, we randomized the assignment to different scenarios, and of the order of the vignettes shown to each participant to reduce the effects of framing due to scenario order. We randomized the order of the three 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 limited surveys to a 20-minute, average time estimate for completion.

External validity concerns how well our results generalize to the population and situations outside the sample used in the study [26]. Our target population is security analysts with varying expertise. We recruited participants using security mailing lists, while specifically recruiting security experts from a research lab specializing in security. Furthermore, we conducted a security knowledge test to measure the extent of security expertise. Sample bias can arise because our participant sample was drawn from only two U.S. institutions, and because our scenarios are limited to only a few factors.

VII. Discussion, Conclusions and Future Work

We now discuss our results in the presence of inter- and intra-expert uncertainty among analysts' security decisions, and we explain the reliability of our security assessment system.

Inter-expert uncertainty is the uncertainty that exists between multiple analysts [16]. Security analysts, in particular, demonstrate this uncertainty by disagreeing on the same scenario [12] or artifact [13]. Our method does not rely on a single analyst's assessment: if the analyst experiences uncertainty, then other analysts' judgments would reduce the uncertainty, unless all analysts are uncertain. As shown in Section V, the two participants P5 and P8 stated that a good SSL/TLS protocol is sufficient to defeat a man-in-the-middle attack, even if the network is public Wi-Fi. While these analysts believe that SSL/TLS is sufficient, others argue that this is insufficient over public Wi-Fi and they recommend using a secure VPN. This is an example of inter-expert uncertainty. To illustrate, if a user is connected over public Wi-Fi, and they are visiting a non-SSL website before being redirected to an SSL-enabled website, then it is easy for a malicious adversary to hijack the session and redirect the user to a website with a forged certificate. Furthermore, the attacker can use certificates signed with trusted certificates, which can cause the SSL connection to appear safe in the browser [10, 14]. Rare events and recent advances in technology illustrate the need for decision-support tools that address limitations of human perception and memory, such as the over- or under-estimation of risk. Cognitive psychologists argue that human memory can fail to recall relevant facts, which can be used to inform decision support models, theories and frameworks to yield intelligent systems [5]. Even the "best" expert can make mistakes and needs support with their evaluation.

Intra-expert uncertainty is the uncertainty that one analyst exhibits in their own judgment [16]. In follow-up interviews, we observed how three experts provided different ratings of the same factors, highlighting variation in opinion over time, e.g., because of new information or a change in context. Other factors that affect intra-expert uncertainty include how representative a scenario appears, or how available the analyst's knowledge of recent events are when passing judgment [28]. In a prior study [12], the SSL Heartbleed vulnerability that affects OpenSSL had recently been announced and this event affected participants' responses about adequacy ratings for SSL [12]. Thus, surveys to collect adequacy ratings may need to be repeated to react to the evolving influences of certain events.

We choose IT2 FSs to build our assessment system because they have the capacity to individually model inter- and intra-expert uncertainties. As shown in our results, we interviewed nine participants in order to verify 36 test scenarios. In only six scenarios (19%), participants disagreed with the security assessment system. In all these six test cases, the security assessment system was more conservative compared to the participants' ratings, i.e., the system indicates inadequate when the participant reports adequate, or the system indicates adequate for a situation that the participant reports as excessive. Participant P9 commented, "in security, I prefer a conservative system's rating like that."

Rule reduction improves readability by human analysts. In Section IV, we show how the rule base is derived from expert-ratings in factorial vignette surveys and we present heuristics to omit unnecessary inputs in the rule antecedents. However, this method has a limitation in that it does not model situations that are absent from the dataset. For example, in the scenarios that we studied, we cannot model requirements combinations that are excessive or adequate overall, because these were not present in survey data. However, this limitation can be addressed by improving the survey design and using expert focus groups aimed at discovering scenarios wherein security is deemed excessive.

Fuzzy logic has been applied in multiple domains [19], including security [7, 15, 25]. Fuzzy data mining techniques using Type-1 Fuzzy Logic have been introduced in intrusion detection systems and have shown an improved outcome [7, 15, 25]. De Ru and Eloff proposed modelling risk analysis using Type-1 Fuzzy Logic and explain that modelling risk analysis with fuzzy logic produces system recommendations that are very close to real situations. They argue that without such systems, organizations run the risk of over- or under-estimating security risks [24]. In this work, we have shown how sometimes analysts underestimate the risk when our assessment system provides more conservative ratings in 19% of the test scenarios. De Ru and Eloff's Type-1 Fuzzy Logic system was not based on security knowledge elicited from multiple experts.

In this paper, we introduced a new approach to build an automated security assessment system based on IT2FLSs. We use survey data collected from 174 security experts to derive the IT2FL rules, and we built membership functions based on this data. Finally, we evaluated the system by running 52 test scenarios on 13 participants. Results indicate that the system succeeds in providing a reliable assessment to analysts, although, it was more conservative in 19% of the 52 scenarios by assessing the security to be lower than our human evaluators.

In future work, we plan to construct scenarios for richer environments based on multi-step attack vectors. In addition, we aim to study ways to recommend to security analysts which requirements will achieve higher overall security ratings. We also plan to complement expert ratings with real-world vulnerability data to assess variability across experts. Finally, we are interested in studying ways to compute the adequacy ratings from requirements class members to help novice analysts learn which security mitigations specifically increase or decrease overall system security based on expert judgments.

Acknowledgment

We thank the study participants, the CMU Requirements Engineering Lab, and Dr. Florian Schaub for their support and feedback. This research was funded by NSA Award #141333 and ONR Award #N00244-16-1-0006.

References

[1] L. Baresi, L. Pasquale, and P. Spoletini, "Fuzzy goals for requirements-driven adaptation," IEEE *18th Int'l Req'ts Engr. Conf.*, pp. 125–134, 2010.

[2] A. Cailliau and A. van Lamsweerde, "Handling knowledge uncertainty in risk-based requirements engineering," *IEEE 23rd Int'l Req'ts Engr. Conf.*, pp. 106-115, 2015.

[3] O. Castillo, P. Melin, and J. R. Castro, "Computational intelligence software for interval type-2 fuzzy logic," *Comput. Appl. Eng. Educ.*, 21(4): 737–747, 2013.

[4] Cisco Systems, Inc., "Cisco 2014 Annual Security Report," Cisco Systems, Inc., 2014.

[5] A. F. Collins, *Theories of Memory*. Psychology Press, 1993.

[6] N. Esfahani and S. Malek, "Uncertainty in self-adaptive software systems," *Soft. Engr. for Self-Adaptive Sys. II*, Springer,, 2013.

[7] G. Florez, S. M. Bridges, and R. B. Vaughn, "An improved algorithm for fuzzy data mining for intrusion detection," in *Annual Meeting of the North American Fuzzy Inf. Processing Society.*, pp. 457–462, 2002.

[8] M. Furr, *Scale construction and psychometrics for social and personality psychology*. SAGE Publications Ltd, 2011.

[9] D. Garlan, "Software engineering in an uncertain world," *FSE/SDP W'shp Future Soft. Engr. Res.*, pp. 125–128, 2010.

[10] W. El-Hajj, "The most recent SSL security attacks: origins, implementation, evaluation, and suggested countermeasures," *Secur. Commun. Netw.*, vol. 5, no. 1, pp. 113–124, 2012.

[11] H. Hibshi and T. D. Breaux, "Evaluation of Lingiustic Labels Used in Applications," Tech. Rep, Carnegie Mellon Uni., 2016.

[12] H. Hibshi, T. Breaux, and S. B. Broomell, "Assessment of Risk Perception in Security Requirements Composition," *IEEE 23rd Int'l Req'ts. Engr. Conf.*, pp. 146-155, Aug. 2015.

[13] H. Hibshi, T. D. Breaux, M. Riaz, and L. Williams, "A Grounded Analysis of Experts' Decision-Making during Security Assessments," To Appear: *Journal of Cybersecurity*, 2016.

[14] L. S. Huang, A. Rice, E. Ellingsen, and C. Jackson, "Analyzing forged ssl certificates in the wild," *IEEE Symp. On Security and privacy (sp)*, pp. 83–97,2014.

[15] J. Luo and S. M. Bridges, "Mining fuzzy association rules and fuzzy frequency episodes for intrusion detection," *Int. J. Intell. Syst.*, vol. 15, no. 8, pp. 687–703, 2000.

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

[17] J. M. Mendel, "Type-2 fuzzy sets and systems: an overview," *IEEE Comput. Intell. Mag.*, 2(1): 20–29, Feb. 2007.

[18] C. I. Mosier, "A critical examination of the concepts of face validity.," *Educ. Psychol. Meas.*, 1947.

[19] J. Mendel and D. Wu, *Perceptual computing: aiding people in making subjective judgments*, v. 13. John Wiley & Sons, 2010.

[20] M. B. Ozek and Z. H. Akpolat, "A software tool: Type-2 fuzzy logic toolbox," *Comp. Appl. Eng. Educ.*, 16(2): 137–146, 2008.

[21] L. Pasquale and P. Spoletini, "Monitoring fuzzy temporal requirements for service compositions: Motivations, challenges and experimental results," *Workshop on Req'ts. Engr. for Sys., Services and Systems-of-Systems (RESS)*, pp. 63–69, 2011.

[22] PricewaterhouseCoopers, "Turnaround and transformation in cybersecurity: Key findings from The Global State of Information Security Survey 2016," 2016.

[23] L. P. Rees, J. K. Deane, T. R. Rakes, and W. H. Baker, "Decision support for Cybersecurity risk planning," *Decis. Support Syst.*, vol. 51, no. 3, pp. 493–505, 2011.

[24] W. G. De Ru and J. H. Eloff, "Risk analysis modelling with the use of fuzzy logic," *Comput. Secur.*, vol. 15, no. 3, 1996.

[25] G. B. Smith and S. M. Bridges, "Fuzzy spatial data mining," *IEEE Trans. Knowl. Data Eng.*, 2002.

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

[27] A. Setalvad, "Demand to fill cybersecurity jobs booming," *Peninsula Press*, 31-Mar-2015.

[28] A. Tversky, D. Kahneman. "Judgment under uncertainty: heuristics and biases." *Science*, 185(4157): 1124-1131, 1974.

[29] U.S. Bureau of Labor Statistics., "Information Security Analysts : Occupational Outlook Handbook:: U.S. Bureau of Labor Statistics." [Online]. [Accessed: 08-Mar-2016].

[30] C. Wagner, "Juzzy-a java based toolkit for type-2 fuzzy logic," *2013 IEEE Symp. on Advances in Type-2 Fuzzy Logic Sys.*, 2013.

[31] C. Wagner, S. Miller, J. M. Garibaldi, D. T. Anderson and T. C. Havens, "From Interval-Valued Data to General Type-2 Fuzzy Sets," in *IEEE Transactions on Fuzzy Systems*, vol. 23, no. 2, pp. 248-269, April 2015.

[32] C. Wagner, M. Pierfitt, and J. McCulloch, "Juzzy online: An online toolkit for the design, implementation, execution and sharing of Type-1 and Type-2 fuzzy logic systems," in *IEEE Int'l Conf. on Fuzzy Sys. (FUZZ-IEEE)*, 2014, pp. 2321–2328.

[33] D. Wu, "A brief Tutorial on Interval type-2 fuzzy sets and systems," *Fuzzy Sets Syst.*, 2010.

[34] D. Wu and J. M. Mendel, "Designing practical interval type-2 fuzzy logic systems made simple," *IEEE Int'l Conf. Fuzzy Sys. (FUZZ-IEEE)*, 2014.

[35] H. Yang, A. D. Roeck, V. Gervasi, A. Willis, and B. Nuseibeh, "Speculative requirements: Automatic detection of uncertainty in natural language requirements," *20th IEEE Int'l Req'ts. Engr. Conf.*, pp. 11–20, 2012.

[36] L. A. Zadeh, "Fuzzy sets," *Inf. Cont.*, 8(3): 338–353, 1965.