Expert Systems

Article

DOI: 10.1111/exsy.12039

Indicators of expert judgement and their significance: an empirical investigation in the area of cyber security

Hannes Holm, Teodor Sommestad, Mathias Ekstedt and Nicholas Honeth

Royal Institute of Technology, SE-10044 Stockholm, Sweden
E-mail: hannesh@ics.kth.se

Abstract: *In situations when data collection through observations is difficult to perform, the use of expert judgement can be justified. A challenge with this approach is, however, to value the credibility of different experts. A natural and state-of-the art approach is to weight the experts' judgements according to their calibration, that is, on the basis of how well their estimates of a studied event agree with actual observations of that event. However, when data collection through observations is difficult to perform, it is often also difficult to estimate the calibration of experts. As a consequence, variables thought to indicate calibration are generally used as a substitute of it in practice. This study evaluates the value of three such indicative variables: consensus, experience and self-proclamation. The significances of these variables are analysed in four surveys covering different domains in cyber security, involving a total of 271 subjects. Results show that consensus is a reasonable indicator of calibration. The mean Pearson correlation between these two variables across the four studies was 0.407. No significant correlations were found between calibration and experience or calibration and self-proclamation. However, as a side result, it was discovered that a subject that perceives itself as more knowledgeable than others likely also is more experienced.*

Keywords: expert judgement, decision support, calibration, cyber security

1. Introduction

Decision-support models are valuable tools when deciding between possible alternatives regarding almost any topic. For such models to accomplish their purpose, their predictions should be both accurate and useful. Ideally, predictions are carried out using a knowledge base built upon large-scale observations of the variables of interest. For instance, if the relation between the number of smokers in a municipality and the number of patients with lung cancer in that municipality is of interest, then data are preferably gained by observing this relation over a sufficient amount of time (e.g. a number of years). However, for various reasons, it is not always feasible to collect data in such a way (Shanteau *et al.*, 2002; Weiss & Shanteau, 2003). For example, the means to gather it can be unavailable or there could be resource constraints preventing it (Gigerenzer & Todd, 1999). In such circumstances, expert judgement can be justified; that is, to elicit knowledge from domain experts on the variable(s) of interest. When expert judgement is used, data quality is uncertain. It is not always the case that an expert is well *calibrated*, that is, the quantities assessed by the expert agree with actual observed values (Cooke, 1991). As multiple experts often are consulted, it is important to determine how much emphasis that should be placed on the data provided by each consulted expert. For example, if a linear pool is used to combine the experts' estimates then the weight of each expert must be determined. Those less calibrated should have less influence on predictions by the resulting decision-support model(s). Various measures of calibration have been proposed for ensuring high data quality when using expert judgement. However, few such measures have been empirically examined and at best only for a few domains (cf. Sections 2.1–2.8).

An increasingly important domain where viable decision support is lacking is the area of cyber security. As there is a great need of quantitative data, there are various constraints prohibiting data collection through observations. One of the more commonly discussed reasons is the potential economical deficits for enterprises sharing cyber security incident information in the public domain (Campbell *et al.*, 2003; Cavusoglu *et al.*, 2004). As a result, there are very little quantitative data available (Verendel, 2009), and many researchers have turned to expert judgement as a viable option (e.g. Madan *et al.*, 2002; Taylor *et al.*, 2002; Haimes, 2003; Holm *et al.*, 2011). Although there are methods available for eliciting data through expert judgement (e.g. Cooke, 1991; Weiss & Shanteau, 2003), most studies in the cyber security area still assign experts equal weight or use simple methods for assigning weights to the subjects.

This study empirically evaluates the performance of three very commonly applied indicators of calibration, namely, experience, self-proclamation and consensus. These variables are evaluated by analysing the results from four studies in the area of cyber security, involving a total of 271 subjects. The studied topics are all highly important issues in the area of cyber security: intrusion detection, denial of service attacks, arbitrary code injection attacks and software vulnerability discovery.

The remainder of the paper is structured as follows: Section 2 describes a literature review and the variables empirically studied in this paper. Section 3 describes the four

studies in terms of elicitation instrument, chosen population and samples. Section 4 details how the studied variables were operationalized in this study. Section 5 describes the results from the study. Sections 6 and 7 discuss the results, and Section 8 critically examines the reliability and validity of research findings. Finally, Section 9 concludes the paper.

2. Measuring the performance of expert judgement

Experts, to be of practical use, should be measurable as improved performance over forecasts or diagnoses given by those people or systems thought of as "inexpert" (Hoenig, 1985).

The validity-based approach, to compare an actual outcome to an expert's assessment, is an appealing measure of the performance of expert judgement because of its simplicity. However, this approach is often impractical as experts are needed in situations where correct answers seldom exist (Gigerenzer & Todd, 1999).

There are various variables that have been proposed for the purpose of reflecting the calibration of experts. This section discusses some of the more significant ones and is concluded with a summary of the variables studied in this paper.

2.1. Self-proclamation

One common way to identify (and weight) experts is through self-proclamation (Abdolmohammadi & Shanteau, 1992; Weiss & Shanteau, 2003); that is, to ask the expert how he or she perceives his or her knowledge in relation to others in the field (Ayyub, 2001). A flaw in this approach is that different subjects can have different egos (Ayyub, 2001), another that being an expert in a field does not necessarily mean that the expert knows anything about the level of knowledge of other people in that field.

2.2. Certification

Shanteau *et al*. (2002) notice that certification, that is, some form of accreditation, often is seen as a reflection of an expert's skill. For example, a university faculty may be 'board certified'. There is, however, a significant problem related to this variable – certification does not always imply skill (Shanteau *et al*., 2002). For instance, people generally move up on the certification ladder but very rarely down. Even if their performance declines, their rank remains.

2.3. Experience

Many studies use the years of job-relevant experience as a surrogate for expert judgement performance (Shanteau *et al*., 2002). The idea is that subjects with more experience of a domain also should perform better in predicting events regarding the domain. There are, however, numerous examples of individuals who never become experts. For instance, Trumbo *et al*. (1962) and Goldberg (1968) have examined the relation between performance and experience. Neither of the studies found any relation between the two variables. Nevertheless, experience is frequently used for valuing expert judgement (Shanteau *et al*., 2002).

2.4. Social acclamation

Another common method for identifying experts is that of social acclamation (Shanteau *et al*., 2002); that is, to ask professionals whom they consider an expert. One such example is that of Phelps (1978) who asked professionals in agriculture whom they considered the best expert (yielding four subjects). A critical flaw to this approach is the 'popularity effect' – an individual better known to its peers is more likely to be seen as an expert (Shanteau *et al*., 2002).

2.5. Consensus

Einhorn (1972, 1974) proposed that consensus, that is, agreement between subjects, is a necessary condition for expertise. If there is disagreement between subjects, then at least some of the would-be experts are not really what they claim to be. Ashton (1985) studied the relation between consensus and calibration for predictions of interest to accountants and found a strong relation between the two variables. Consensus has been used to compare the performance of different expert judgement techniques on many occasions, for example (Fischer, 1981). There are, however, critics of consensus. In particular, Shanteau (2001) and Weiss and Shanteau (2003) argue that consensus isn't an appropriate criterion for expertise. Constructs, such as the defining characteristics of a disease, must be shared by the linguistic community that employs them (Weiss & Shanteau, 2003). For example, there is a need for agreement regarding what is meant by the term glaucoma. The capability of (and process for) identifying symptoms, however, depend on the examiner's perceptual and integrative skills. Thus, different doctors can make different predictions, for example, whether a patient has glaucoma or not. To sum up, many experts may agree – but they may all be wrong (Shanteau, 2001; Weiss & Shanteau, 2003).

2.6. Creation of experts

In some areas, it might be possible to give subjects extensive training, thus in a sense creating experts (Shanteau *et al*., 2002). For example, Chase and Ericsson (1981) trained a student to increase his short-term memory so that he managed to set a new world record on the topic. However, for obvious reasons, this procedure is not suitable for all contexts.

2.7. Knowledge tests

One way to measure the relative knowledge of a subject is to ask the subjects questions for which the answer is known beforehand, or will be known before the analysis is carried out. The subjects' performance on these questions can then be used to identify whether they are experts or not. An established method using knowledge tests for weighting the judgement of experts is the classical model developed by Cooke (1991). The performance of this model has previously been evaluated (Cooke, 2008) and found to outperform both equally weighted experts' judgement and the judgement by the best expert. A problem with knowledge tests is, however, that it can be difficult to elicit which facts to apply in a given situation – especially as expert opinion is needed in domains where correct answers seldom exist (Gigerenzer & Todd, 1999). Also, asking subjects additional questions could cause

reliability issues due to the survey or interview ending up as too extensive (Janes, 1999).

2.8. Cochran–Weiss–Shanteau

Weiss and Shanteau (2003) propose usage of a variable they name Cochran–Weiss–Shanteau, or CWS. The authors argue that two necessary characteristics of expertise are discrimination of the various stimuli in the domain and consistent treatment of similar stimuli. These two variables construct the foundation of the variable proposed by the authors (Weiss & Shanteau, 2003). Unfortunately, a high CWS does not necessarily imply expertise. For example, a doctor who administrates treatments primarily based on patients' hair colour would perform well according to the CWS as long as hair colour is discriminated consistently among the treated patients. Furthermore, the variable requires additional questions in the survey to measure inconsistency, something that could cause reliability issues due to the survey ending up as too extensive (Janes, 1999).

2.9. Studied variables

All the variables discussed in Section 2 come with various strengths and weaknesses. This study analyses the value of three very commonly used indicators of calibration, namely, experience, self-proclamation and consensus. Calibration is operationalized in the same fashion as by Cooke (1991); that is, as knowledge tests measuring the extent to which the quantities assessed by that expert agree with actual observed values. A total of 41 different test questions believed to be representative to the domains in question are employed for this purpose. The operationalization is detailed in Section 4.1, and the validity of it is discussed in Section 8.

There are few studies that have made any empirical comparisons of the relative performance of consensus, experience and self-proclamation (cf. Sections 2.1–2.8). Also, there is, to the authors' knowledge, not a single study of these variables that has been carried out in the area of cyber security. This study analyses the performance of consensus, experience and self-proclamation in the context of four domains in the cyber security field. The operationalizations of these variables during the present study can be found in Section 4.2 (Consensus), Section 4.3 (Experience) and Section 4.4 (Self-proclamation). The primary topic of interest is formulated in the first research question (RQ) of the study, described as follows.

RQ1 Are experience, self-proclamation or consensus related to calibration?

This research question can be answered through three hypotheses, which are defined as follows.

- H1: There is a positive correlation between experience and calibration.
- H2: There is a positive correlation between self-proclamation and calibration.
- H3: There is a positive correlation between consensus and calibration.

In addition to correlations with calibration, experts are commonly thought to share many characteristics (Shanteau, 1988; Shanteau, 2001; Farrington-Darby & Wilson, 2006). Thus, an expert that perceives himself or herself as more knowledgeable should also have a higher consensus with others (Shanteau, 1988; Farrington-Darby & Wilson, 2006) and possess more experience (Shanteau, 1988; Farrington-Darby & Wilson, 2006). Also, an expert that displays a high consensus with other experts should have a high level of experience (Shanteau, 2001). This is described through the second research question of the study.

RQ2 Which correlations exist between experience, self-proclamation and consensus?

The second research question can, in the same sense as the first, be answered through three hypotheses.

- H4: There is a positive correlation between experience and self-proclamation.
- H5: There is a positive correlation between experience and consensus.
- H6: There is a positive correlation between consensus and self-proclamation.

Figure 1 illustrates the studied variables and their hypothesized relations during this project. The primary hypotheses of the study are shown with straight lines (H1–H3), and the secondary hypotheses of the study are shown with dashed lines (H4–H6).

3. Expert judgement in four studies

Computer and network security, or cyber security, are critical issues (Bishop, 2003). The importance of cyber security is steadily increasing in relation to the development of wide spread global infrastructure technologies and progressively more complex enterprise information technology environments (Hansman & Hunt, 2005). Organizations are forced to spend large amounts of resources on cyber security matters because successful cyber attacks can be extremely expensive. Sound decision-support models for cyber security would enable decision makers to take more well-informed choices, for example, when choosing between different security protection mechanisms. Although there are numerous approaches to measuring cyber security (e.g. Humphreys, 2006; Den Braber *et al.*, 2007; Sommestad *et al.*, 2010), only a handful have been tested with respect to validity (Verendel, 2009). As a result, there is no 'golden standard' for assessing cyber security. One reason behind this is the lack of quantitative data – most enterprises do not want to share their cyber security incidents. One important reason is the cost of

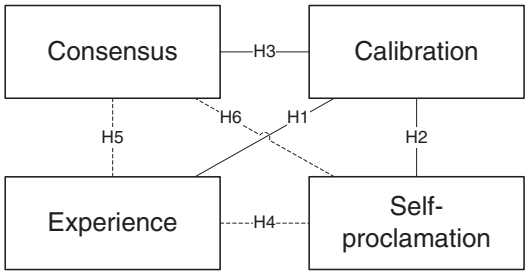

Figure 1: *Studied variables and their hypothesized relations.*

doing so (Campbell *et al.*, 2003; CavSusoglu *et al.*, 2004). It is also a field where measurement can be difficult. For instance, it is virtually impossible to assert that confidential business data have not been read by unauthorized individuals. Cyber security is thus an area for which expert judgement certainly can be justified and various researchers have turned to it as a viable option (e.g. Madan *et al.*, 2002; Taylor *et al.*, 2002; Haimes, 2003; Holm *et al.*, 2011).

There are various domains in the area of cyber security, for example, vulnerability discovery (Alhazmi *et al.*, 2007), social engineering (Dodge, 2007), denial of service (Hansman & Hunt, 2005), code injection (Hansman & Hunt, 2005) and intrusion detection (Axelsson, 2000). Taxonomies that describe the field can be found in the works of Laprie *et al.* (2004) and Hansman and Hunt (2005). This project studies four of the more important domains: intrusion detection systems (Study A) (Sommestad *et al.*,), denial of service attacks (Study B) (Sommestad *et al.*, 2011), arbitrary code execution attacks (Study C) (Sommestad *et al.*, 2012a) and software vulnerability discovery (Study D) (Sommestad *et al.*, 2012b). As discussing these matters is out of scope for the paper, the interested reader is referred to the references in this section for further information.

3.1. Population and sampling

The four conducted studies aimed to identify quantities related to cyber security. Thus, the subjects needed both the ability to evaluate aspects in the domains and the ability to reason in terms of probabilities. In terms of the expert categories described by Weiss and Shanteau (2003), individuals that are expert judges are desirable. Studies of experts' calibration have concluded that experts are well calibrated in situations with learnability and with ecological validity (Bolger & Wright, 1994). Learnability comes with models over the domain, the possibility to express judgement in a coherent quantifiable manner and the opportunity to learn to from historic predictions and outcomes. Ecological validity is present if the expert is used to making judgements of the type they are asked for.

Researchers in the cyber security field have performed and disseminated a number of empirical studies related to effectiveness of different solutions. Although these studies sometimes are questionable with respect to generality (McHugh, 2000), they do offer input to specific scenarios. A practitioner (e.g. a system operator) will probably not have the same opportunity to learn the effect of different scenarios because they typically only have experience from a few installations and rarely perform stringent evaluations of effectiveness. Also, with respect to ecological validity, it was expected that researchers would be more used to estimating probability distribution and reason in terms of probabilities.

Cyber security researchers were defined as the population for all four studies. These can be expected to both understand how to reason with probabilities and also possess the required skills to evaluate the effectiveness of different solutions. They also perform experiments more frequently than practitioners and thus have a better possibility to build knowledge regarding the domain.

To identify suitable subjects, articles published in the SCOPUS database (Elsevier, 2010), Inspec (Elsevier Inc., 2010) and Compendex (Elsevier Inc., 2010) between January 2005 and September 2010 were reviewed. Authors who had written articles in the information technology field with keywords in the title, abstract or keywords, which were related to the domain in question, were identified.

If their contact information could be found, they were added to the list of *potential subjects*. After reviewing and screening subjects and their contact information, remove subjects were *invited to a web survey*. Some of these invited subjects had outdated or incorrect contact information, resulting in only some *mails reaching their destination*. Furthermore, only a subset of all mails reaching their destination had a *survey opened by the subject*. Of all subjects that opened the surveys, not all *submitted answers*, and only a subset of these were *complete* (cf. Table 1).

3.2. Construction of elicitation instruments

Web surveys were used to query subjects on the importance of different variables in the four studied domains. The number of questions in each survey can be found in Table 1. Each surveyed variable (i.e. survey question) was identified through a literature review in combination with validating interviews, as recommended by Gable (1994). Every survey comprised four parts, each beginning with a short introduction to the section. First, the subjects were given an introduction to the survey that explained the purpose of the survey and its outline. In this introduction, they also confirmed that they were the person who had been invited and provided information about themselves, for example, years of experience in the field of research. Second, the subjects received training regarding the answering format used in the survey. After confirming that this format

Table 1: *Information regarding the carried out surveys*

	Pilot survey**	Study A	Study B	Study C	Study D
Survey questions	11*	8	11	11	11
Validating interviews	4		1	2	2
Potential subjects	13561		1378	964	2211
Invited to web survey	500	5769	1065	545	384
Mail reaching their destination	373	4200	885	445	300
Survey opened by subject	123	1355	296	119	92
Submitted answer	34	243	65	22	17
Complete answer	34	165	35	21	16

*Three added versions of one question to measure the internal consistency of the survey question format.

**The pilot survey was carried out using the survey of Study A, using 500 of the total 6269 identified subjects (chosen through simple random sampling).

was understood, the subjects proceeded to its third part. In the third part, the questions of the study were presented to the subjects. Finally, the subjects were asked to provide qualitative feedback on the survey. As recommended by Cavusgil and Elvey-Kirk (1998), motivators were presented to the subjects invited to the survey: (1) helping the research community as whole; (2) the possibility to win a gift certificate on literature; and (3) being able to compare their answers to answers by other experts after the survey was completed.

For each survey question, the subjects were asked to provide a probability distribution that expressed their beliefs regarding the likely distribution of the answer. In the survey, the subjects specified their distributions by adjusting sliders or entering values to draw dynamically updated graphs over their probability distributions. The three points specified by the subjects [the 5th percentile, the 50th percentile (the median) and the 95th percentile of the probability distribution] were entered. These define four intervals over the range [0%, 100%]. The graphs displayed the probability density as a histogram, instantly updated upon change of the input values. Use of graphical formats is known to improve the reliability of elicitation (Garthwaite *et al*., 2005). Figures and colours were also used to complement the textual questions and make the questions easier to understand.

Elicitation of probability distributions is associated with a number of issues (Garthwaite *et al*., 2005). Effort was therefore spent on ensuring that the measurement instrument held sufficient quality. The surveys were after careful construction qualitatively reviewed during personal sessions with several external subjects representative of the domains (cf. Table 1). These sessions contained two parts. First, the subjects were given a task to fill in the survey, given the same amount of information as someone doing it remotely. After this, discussions followed regarding the instrument quality. These sessions resulted in several improvements.

Another part of the instrument review concerned the internal validity of the question format as such: a pilot study using a simple randomized sample of 500 subjects of the 6269 invited subjects of Study A (cf. Table 1). This pilot survey was opened by 123 persons and completed by 34 during the week it was open. Cronbach's alpha (Cronbach, 1951; Cronbach & Shavelson, 2004) is often used to test the reliability of a survey instrument and if subjects understand its questions. A reliability test using Cronbach's alpha was carried out using one variable (four different versions of one of the survey's questions). Measuring the reliability of more than one question would be inefficient, as all sections and questions were formatted in the same way and most likely have created bias for the instrument used during the pilot study. Results from this test showed a Cronbach's alpha of 0.817, which indicates good internal consistency of the instrument. Also, qualitative comments entered in the survey's feedback section confirmed that subjects understood the questions.

4. Operationalizations of the studied variables

This section describes how the variables of interest were measured. This study uses the same method for correlating calibration, consensus, experience and self-proclamation as Ashton (1985) used to correlate calibration and consensus; that is, Bivariate Pearson correlation analysis (Warner, 2008). Two-tailed hypotheses tests (*t*-tests) (Warner, 2008) are used to study the hypotheses stated in Section 2.9. The null hypothesis of the hypotheses tests, H0, is that correlation coefficient between the two tested random variables is zero. The boundary associated with rejecting a null hypothesis is generally described using probability, p. A commonly used level of significance is $\alpha = 0.05$, so that if $p < 0.05$, then it implies that the null hypothesis shall be rejected (Warner, 2008). On the other hand, if $p \geq 0.05$, the stated hypotheses H1 to H6 shall be rejected. This level of significance is employed when analysing the results from this study.

4.1. Operationalization of calibration

Calibration concerns the extent to which assessed quantities by an expert agree with observed values (Cooke, 1991). Calibration in the four areas was measured using survey questions for which the answer was already known (cf. Section 3.2), a typical method of operationalizing the variable (Ashton, 1985; Cooke, 2008). It is important that the surveyed questions are representative to the scope of the study. How this was handled is described in Section 8. A total of 41 questions were used; the complete set of questions can be found in Appendix E.

There exist several different methods that can be applied to measure the calibration of a subject, for example, entropy and Euclidian distance (Cooke, 1991; Ayyub, 2001). A very well-established method for measuring the calibration of a set of experts is the model developed by Cooke (1991), and this is also what was used during this study. The remainder of this section describes how calibration through Cooke's method is measured.

As the answer to each question used for performance evaluation is an uncertain quantity to the experts, they are asked to specify a probability distribution that represents their belief about its true value. This distribution is typically specified by stating its 5th, 50th and 95th percentile values; this was also the case for the four studies analysed in this paper. The three percentiles yield four intervals over the percentiles [0–5, 5–50, 50–95, 95–100] with probabilities of $p = [0.05, 0.45, 0.45, 0.05]$. As the true values of the questions are realizations of these variables, the well calibrated expert will have approximately 5% of the realizations in the first interval, 45% of the realizations in the second interval, 45% of the realizations in the third interval and 5% of the realizations in the fourth interval (Cooke, 1991).

When true values and expert estimates are known, the calibration score can be calculated for each subject. If s is the distribution of the seed over the intervals, the relative information of s with respect to p is

$$I(s,p) = \sum_{i=1}^{4} \mathrm{In}(s_i/p_i)$$

Loosely speaking, this value indicates how surprised someone would be if one believed that the distribution was p (the expert's assessment) and then learnt that it was s (*the actual value*). If N is the number of samples (seeds), the statistic of $2NI(s, p)$ is asymptotically chi-square distributed with three degrees of freedom. When continuous variables with three quantiles are elicited (such as during the

present study), 8–10 seed questions are sufficient to accept the chi-square approximation (Cooke, 1991). This asymptotic behaviour is used to calculate the calibration *Cal* of expert *e* as

$$Cal(e) = 1\text{-}\chi^2_{(3)}(2N\ \mathrm{I}(s,p))$$

This calibration measures the statistical likelihood (*p*-value) of the hypothesis that realizations of the real values (*s*) are sampled independently from distributions agreeing with the expert's assessments (*p*) (Cooke, 1991). In other words, *Cal* (*e*) is the probability that it would be incorrect to regard expert (*e*) as not calibrated on the basis of the observations made – the risk of being incorrect is low if an expert (*e*) with a low *Cal*(*e*) is categorized as not calibrated and high if *Cal*(*e*) is high.

4.2. Operationalization of consensus

This study uses the same method to measure consensus as those of Ashton (1985); that is, pairwise correlational consensus (Ashton, 1985). The pairwise correlational consensus is based on the consensus measure traditionally discussed in papers, that is, the correlation between the predictions of each *pair* of subjects. For example, in the case of the 16 experts in Study D, the pairwise correlational consensus consist of the Pearson correlation coefficient (Warner, 2008) between the 33 survey answers (11 answers, each including 3 percentiles) made by each pair of the 16 subjects. Table 2 shows the pairwise correlational consensus of the 16 experts in Study D. It does for example show that the correlation between the answers by subject 1 and subject 2 is 0.732. The individual subjects' mean consensus were then calculated using Fisher's *z*-transformation (Glass & Stanley, 1970).

The consensus matrices for the other analysed studies are not included in the paper because of the space they require. They can, however, be downloaded from[1] or provided upon request to the authors.

4.3. Operationalization of experience

Experience was measured, as is commonly practised (Shanteau *et al.*, 2002), on a scale of years. The exact formulation of the question in the survey was 'Please specify how many years you have been doing research related to [DOMAIN]', where [DOMAIN] corresponds to the domain in question (e.g. 'software vulnerabilities and software exploits').

4.4. Operationalization of self-proclamation

Self-proclamation was measured on a scale from 0% to 100% where the specified value corresponded to how knowledgeable the subject perceived himself or herself compared with others in the same field of research. For example, 'Top 90%' means that the subject perceived himself or herself to be more knowledgeable than 10% of

[1]www.ics.kth.se/expert_consensus.xls

Table 2: *Pairwise correlational consensus scores of Study D*

Subject	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
1	–															
2	0.732	–														
3	0.718	0.864	–													
4	0.8	0.815	0.887	–												
5	0.329	0.553	0.665	0.642	–											
6	0.656	0.499	0.468	0.398	0.003	–										
7	0.683	0.425	0.38	0.35	0.069	0.738	–									
8	0.109	0.084	0.164	0.382	0.39	−0.149	−0.21	–								
9	0.429	0.57	0.659	0.583	0.195	0.354	0.097	0.127	–							
10	0.577	0.736	0.745	0.701	0.511	0.278	0.052	0.076	0.605	–						
11	0.628	0.792	0.893	0.873	0.707	0.244	0.273	0.412	0.617	0.644	–					
12	0.573	0.425	0.36	0.497	0.247	0.315	0.672	0.335	0.053	−0.039	0.467	–				
13	0.289	0.062	0.053	0.274	−0.267	0.002	−0.135	0.472	0.266	0.091	0.146	0.188	–			
14	0.746	0.623	0.826	0.823	0.609	0.443	0.423	0.33	0.578	0.686	0.776	0.35	0.167	–		
15	0.68	0.758	0.722	0.722	0.61	0.545	0.464	0.313	0.465	0.58	0.69	0.521	0.007	0.694	–	
16	0.783	0.689	0.792	0.861	0.473	0.454	0.418	0.46	0.609	0.513	0.825	0.562	0.372	0.795	0.704	–
Mean*	0.611	0.675	0.736	0.755	0.462	0.386	0.365	0.221	0.437	0.481	0.645	0.394	0.156	0.683	0.646	0.651

*Individual Correlational Consensus. Fisher's *z*-transformation (Glass & Stanley, 1970) was used for calculating mean correlations.

the scientific community of the domain in question. The exact formulation of the question was

> "Where would you place yourself compared to other authors of scientific publications related to [DOMAIN] when it comes to general knowledge in the field? (For example, top 20% means that you think that you are more knowledgeable than 80% of this scientific community.)", where [DOMAIN] corresponds to the domain in question (e.g. "software vulnerabilities and software exploits").

The subjects' self-proclamation were transformed from 'Top' to 'Bottom' when comparing this variable to the other variables to create more pedagogical results (e.g. a subject specifying 'Top 70%' had their value translated to 'Bottom 30%'). This was carried out to make the result more pedagogical in the sense that a positive correlation's direction for this variable would mean that it is a good indicator (as for the other indicators).

5. Results from the four studies

Descriptive statistics (arithmetic means and variances) regarding the four studies can be seen in Table 3. There are some differences between the studies: (1) the subjects in Study D are the most calibrated; (2) the subjects tended to agree the most in Study A; and (3) subjects in Study C and D have a higher experience and self-proclamation than those in Study A and Study B. The complete datasets regarding the four variables can be found in Appendix A (Study A), Appendix B (Study B), Appendix C (Study C) and Appendix D (Study D).

Bolger and Wright (1994) argue that an expert's performance is good when ecological validity and learnability are high and poor when ecological validity and learnability are low; that is, the calibration of experts is influenced by the availability of accurate, relevant and objective data and/or domain models upon which decisions can be based. As there are few cyber security models that have been properly validated (Verendel, 2009), it is not surprising that calibration is low for all four studies. The best decision-support models can arguably be found in the domain of Study D, which is consistent with the argumentation by Bolger and Wright (1994).

6. Performance of indicators of calibration

This chapter is categorized in three sections, discussing the hypotheses of the first research question of the study, 'Is experience, self-proclamation or consensus related to calibration?' Calibration is domain-specific (Chi, 1988) and depends on the questions asked. As the analysed studies have different (and various amounts of) questions, it is unfortunately not possible to aggregate the results from the four studies to a single dataset. As a consequence, each hypothesis has four different evaluations, one for each study.

6.1. Experience and calibration

The first hypothesis states that there is a positive correlation between experience and calibration (H1, cf. Table 4); that is, do more years of domain experience imply greater calibration? Of the four analysed studies, the results are rather dissimilar: the result from Study A is significant and indicates that a subject with more years of experience actually is less calibrated in terms of predicting actual observations in the domain; Study B and Study C show positive yet insignificant results; Study D shows the strongest correlation, yet insignificant according to the stated level of significance. However, because of the size of the correlation and the marginal significance, we choose to reject the null hypothesis for Study D (according to a significance level of 10%).

One potential explanation for the varied results could be that in some fields, an increased amount of experience means that an individual actually has less time to perform empirical studies of different properties. In other words, as less experienced individuals can spend effort to empirically study specific properties, effort by experienced individuals could be required for wider matters, for example, to coordinate effort spent by less experienced individuals. This also applies to staying updated with recent advances in a domain: a senior is oftentimes required to spend effort within several different domains (e.g. both vulnerability research and intrusion detection research) rather than, as a junior, within a single domain. Given such a scenario, it can be expected that the actual calibration regarding recent

Table 4: *Experience and calibration*

Study	Correlation coefficient	p	Hypothesis rejected	Samples
Study A	−0.171	0.029	Yes[a]	163
Study B	0.009	0.962	Yes	33
Study C	0.077	0.754	Yes	19
Study D	0.435	0.092	No[b]	16

[a]Correlation not in the hypothesized direction.
[b]Given a 10% significance level.

Table 3: *Descriptive statistics of the four studies*

		Study A	Study B	Study C	Study D
Calibration	Mean	0.157	0.060	0.004	0.300
	Variance	0.044	0.012	0.000	0.041
Experience	Mean	5.640	5.879	6.658	7.750
	Variance	10.888	7.547	26.835	9.800
Self-proclamation	Mean	51.585	43.171	53.263	56.063
	Variance	377.054	515.793	674.316	618.463
Consensus	Mean*	0.636	0.476	0.561	0.541
	Variance*	0.101	0.0517	0.0348	0.064

*Fisher's z-transformation (Glass & Stanley, 1970) was used to calculate means and variances for consensus.

empirical updates within a particular domain is lower than that of an inexperienced individual whose efforts are dedicated towards that single domain.

To conclude, our results suggest that experience on a matter does not assure expertise of that matter. Whether experience is a useful indicator of calibration depends on the context of the studied problem and for which situation it is useful is a topic for further work.

6.2. *Self-proclamation and calibration*

The second hypothesis is that self-proclamation and calibration are positively associated (H2). There were consistent weak negative correlations between calibration and self-proclamation for all four studies (cf. Table 5). These results suggest that a subject who perceives himself or herself as more knowledgeable than others actually possesses less knowledge. However, as none of these correlations are significant, it is most likely a product of random variation. This study did not find any evidence that can justify not rejecting the second hypotheses for any of the four studies, suggesting that it is not reliable to employ self-proclamation as a measure of calibration.

6.3. *Consensus and calibration*

The third hypothesis is that consensus and calibration are positively associated (H3). Consensus showed significant and strong correlations to calibration in two out of four evaluated studies (Study A and Study B), and fairly strong correlations in Study C and Study D, yielding a mean correlation of 0.407 (cf. Table 6); that is, the third hypothesis is not rejected in two out of four cases. In the other two cases, the correlation coefficients are positive, and it is possible that their sample sizes (19 and 16) are too small to produce significant correlations (as suggested by results from Study A and Study B). Future studies, with larger samples, would make it possible to investigate if the sample size is the reason for the insignificant correlations in Study C and Study D.

Table 5: *Self-proclamation and calibration*

Study	Correlation coefficient	p	Hypothesis rejected	Samples
Study A	−0.125	0.110	Yes	164
Study B	−0.043	0.807	Yes	35
Study C	−0.055	0.822	Yes	19
Study D	−0.030	0.912	Yes	16

Table 6: *Consensus and calibration*

Study	Correlation coefficient	p	Hypothesis rejected	Samples
Study A	0.555	<0.0001	No	164
Study B	0.654	<0.0001	No	35
Study C	0.187	0.442	Yes	19
Study D	0.235	0.380	Yes	16

7. Experience, self-proclamation and consensus

This chapter discuss the second research question, 'Which correlations exist between experience, self-proclamation and consensus?' As calibration is not part of this research question, it is possible to aggregate all four studies into a single dataset and analyse the results accordingly.

7.1. *Experience and self-proclamation*

The fourth hypothesis (H4) is that there is a positive correlation between experience and self-proclamation (H4). As the results show a significant positive correlation between experience and self-proclamation (cf. Table 7), there is empirical support for not rejecting the fourth hypothesis. In other words, a subject who has more years of experience likely also perceives itself as having more knowledge than others in the same field of research.

7.2. *Experience and consensus*

The fifth hypothesis is that experience and consensus are positively associated (H5). There is a very weak, non-significant, negative correlation between experience and consensus (cf. Table 8). The fifth hypothesis is thus rejected on the basis of the results gained in this study; that is, experience and consensus are not positively associated.

7.3. *Consensus and self-proclamation*

The sixth hypothesis is that there is a positive correlation between consensus and self-proclamation. As for the fifth hypothesis, there is a weak non-significant correlation between consensus and self-proclamation (H6, cf. Table 9). The results indicate that the sixth hypothesis should be rejected. In other words, the results signify that there is no positive correlation between consensus and self-proclamation.

Table 7: *Experience and self-proclamation*

Correlation coefficient	p	Hypothesis rejected	Samples
0.292	<0.0001	No	231

Table 8: *Experience and consensus*

Correlation coefficient	p	Hypothesis rejected	Samples
−0.019	0.771	Yes	234

Table 9: *Consensus and self-proclamation*

Correlation coefficient	p	Hypothesis rejected	Samples
−0.051	0.439	Yes	234

8. Limitations of the study

This chapter critically discusses the research findings in terms of validity and reliability.

8.1. Validity of the survey questions

In this study, 8–11 questions (depending on the study) were used to evaluate the performance of experts for each study (as recommended by Cooke (1991)). The quality of the results regarding calibration is naturally determined by the quality of these questions. If they are chosen from a very narrow part of a domain, it is unlikely that they can be related to the general competence of the experts. From a critical view, this is a major issue towards the validity of research findings; the results regarding RQ1 can only be viewed from the perspective of the questions used to measure calibration. This study handles this issue through drawing questions from areas strongly related to the domains at issue. Furthermore, interviews with subjects' representative of the queried domains (cf. Table 1) were used to examine the validity of the chosen questions. These interviews did not show any issues regarding the appropriateness of the questions.

8.2. Calibration and secondary data

A threat to the reliability and validity of the results is that the data, which the test questions are based on, were publicly available. Although the authors of all articles used for designing the test questions were excluded from the list of potential subjects, other subjects could potentially have used these data to identify the correct answers. However, it appears unlikely that any of the subjects had done so. None of the subjects answering the survey gave comments that indicate that they had realized that the correct answer could be found this way. The reviewers of the surveys did not perceive this as a likely issue either. Furthermore, inspections of the received answers did not indicate any answers based on these sources.

8.3. Response frequencies of the surveys

Of all the subjects that were invited, a mean of 31% opened the surveys, and a mean of 19% of those who opened a survey completed it. One reason for some not opening the survey is that many invited subjects likely do not use those email accounts anymore. Another important reason for the response rate is likely that the survey was fairly complex, time-demanding and spread over the internet. The mean of 19% completion is reasonable considering the circumstances.

8.4. Reliability of the elicitation instrument

Groves *et al*. (2009), Janes (1999) and Gable (1994) provide guidelines for eliciting data using surveys in general. Cooke (1991) discusses elicitation of data from experts in particular. How these have been addressed in the present study is described later. All these authors state that questions must be clear and unambiguous and that a dry run should be carried out before the actual study. The clarity of questions was tested both through qualitative reviews and dry runs. Strategically selected subjects' representative of the population and domains helped improve the understandability of the instruments. Also, a quantitative test of the surveys' quality was performed in a pilot study with 500 subjects from the 6269 elicited subjects of Study A (selected using a simple randomization technique).

It is also suggested that an attractive graphical format and a brief explanation of the elicitation format should be prepared (Cooke, 1991; Janes, 1999; Groves *et al*., 2009) – especially when probability distributions are used (Garthwaite *et al*., 2005). Both the questions and the answering format used in this study were supported by graphical illustrations. The questions were described in text complemented with figures of the scenarios; the answers were given by entering a probability density function on the screen. Also, background information introduced each new section. This format was also carefully explained in an introductory training section in the survey.

Janes (1999) and Groves *et al*. (2009) argue that the survey should be short. Cooke (1991) makes this statement more tangible and recommends that the elicitation should not exceed 1 h and that coaching should be avoided. None of the subjects who completed the survey spent more than 1 h to do so (the mean time was 23 min), and efforts were made to ensure that the questions were formulated in a neutral way.

The last recommendation given by Cooke (1991) is that an analyst should be present when subjects answer the questions. With a web survey, this was obviously not fulfilled. The subjects were given contact information to the research group when invited to the survey that they were encouraged to use any if questions arose. As this ensures that no coaching occurred during the elicitation, it is possible that it suppressed potential questions being asked. To identify potential issues of this type, the survey subjects were asked to comment the clarity of the questions and the question format used. On the basis of the comment received, no distressing issues relating to the formulations of the questions arose. Also, the quantitative analysis of the reliability of the survey instrument using Cronbach's alpha suggests that its reliability is high (cf. Section 3.2).

9. Conclusions and future work

This study evaluated the performance of three commonly used indicators of calibration: experience, self-proclamation and consensus. The variables were studied in the context of four areas of cyber security: intrusion detection, denial of service attacks, arbitrary code execution attacks and software vulnerability discovery. A total 271 subjects were part of the study. Mean correlations for the six hypotheses in the four analysed studies can be found in Figure 2.

There are several implications of the results. The collected evidence point towards neither experience nor self-proclamation as good indicators of calibration; that is,

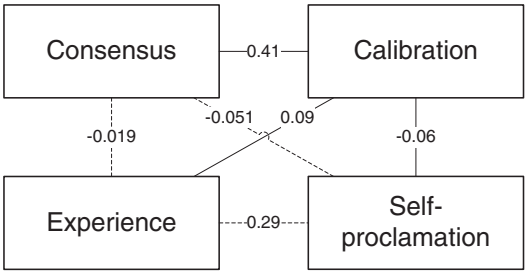

Figure 2: *Mean correlations between the studied variables across the four studies.*

using these variables for choosing experts or weighting their judgement is not a viable approach. Of these two variables, experience displayed interesting results. There were both a significant negative correlation and a strong positive correlation between experience and calibration, suggesting that additional years' experience can both decrease and increase the calibration of an expert. One potential explanation for these results could be that in some fields, an increased amount of experience means that an individual actually has less time to perform empirical studies of different properties and remain updated on recent advances. This relation should, however, be further studied to fully identify the reasons behind these curious results.

In literature, there are several authors that have argued that consensus does not need to suggest calibration, for example (Shanteau, 2001; Weiss & Shanteau, 2003). The results from the present study, however, point towards consensus as a reasonable measure when sorting out, or weighting the answers of, potential experts in a domain. The fact that consensus does not require any tempering with the elicitation tool (e.g. any additional survey questions) makes it an especially attractive approach. It should be noted that the relation between consensus and calibration only has been tested for a few domains. It was tested for four domains of cyber security in this study and demonstrated significant correlations in the hypothesised direction in the two studies with larger samples. It has previously been tested in the accounting domain by Ashton (1985) with signifcant results. It would be valuable to gather additional data regarding this hypothesis, not only in the domains of the present study but also for other domains.

Another interesting result, yet perhaps less of a surprise, is the relation between experience and self-proclamation. This study shows that a subject with more experience perceives itself as more knowledgeable than others. In turn, this hints towards a significant relation between experience and ego that could be interesting to further study.

This study only elicited information from those that had published academic articles. Although it is a very attractive target for expert judgement studies, not all experts publish such articles. It would be useful to gather data from other subjects who are perceived as experts. For example, a very promising role in the area of cyber security is the professional penetration tester – an individual who audits cyber security aspects for a living. It would be valuable to analyse if the results from the present study are contingent to those gained through studying professional penetration testers.

This study also highlights issues generated because of the lack of significant decision-support model(s) in the area of cyber security– the calibration was in general low for all studied topics. It is clear that quantitative data collection would greatly benefit the area, both for research and in industry.

Appendix A: Measurements for Study A

Subject	Consensus	Calibration	Years' experience	Self-proclamation	Subject	Consensus	Calibration	Years' experience	Self-proclamation
1	0.709	0.664	6	90	68	0.610	0.145	5	67
2	0.698	0.664	2	50	69	0.656	0.088	6	40
3	0.697	0.664	8	60	70	0.354	0.088	2	80
4	0.518	0.664	1.5	31	71	0.495	0.088	1	98
5	0.712	0.664	3	70	72	0.785	0.088	3	40
6	0.741	0.664	2	24	73	0.634	0.088	4	80
7	0.706	0.664	2	90	74	0.686	0.080	6	50
8	0.759	0.664	5	40	75	0.437	0.067	2	95
9	0.744	0.664	2	90	76	0.483	0.067	5	40
10	0.748	0.664	1	80	77	0.590	0.067	7	30
11	0.804	0.664	2	50	78	0.718	0.043	7	50
12	0.815	0.664	4	50	79	0.597	0.043	6	50
13	0.801	0.640	3	50	80	0.683	0.034		50
14	0.646	0.541	5	50	81	0.760	0.034	8	10

15	0.792	0.541	6	50	82	0.686	0.034	4	75
16	0.790	0.541	5	50	83	0.778	0.034	10	50
17	0.774	0.534	5	50	84	0.757	0.027	5	40
18	0.832	0.534	8	50	85	0.392	0.027	3	70
19	0.696	0.429	5	40	86	0.457	0.016	6	35
20	0.791	0.429	5	70	87	0.619	0.016	2	60
21	0.748	0.429	4	40	88	0.580	0.016	7	50
22	0.763	0.429	5	55	89	0.667	0.016	5	50
23	0.795	0.429	3	75	90	0.721	0.011	8	70
24	0.766	0.429	20	50	91	0.540	0.009	10	50
25	0.794	0.429	4	48	92	0.598	0.006	2.5	60
26	0.795	0.429	2	25	93	0.648	0.006	10	69
27	0.695	0.429	7	60	94	0.562	0.003	4	40
28	0.760	0.429	10	80	95	0.130	0.003	4	50
29	0.759	0.429	3	50	96	0.633	0.003	9	50
30	0.815	0.286	5	50	97	0.552	0.003	3	70
31	0.684	0.286	4	50	98	0.452	0.003	7	60
32	0.842	0.286	10	50	99	0.646	0.003	3	50
33	0.878	0.286	2	70	100	0.633	0.003	10	25
34	0.859	0.286	5	50	101	0.618	0.003	3	60
35	0.856	0.286	10	10	102	0.352	0.003	10	10
36	0.876	0.286	4	50	103	0.295	0.002	10	35
37	0.836	0.286	4	30	104	0.181	0.002	2	75
38	0.867	0.286	4	50	105	0.533	0.002	3	25
39	0.819	0.286	4	70	106	0.372	0.002	8	50
40	0.822	0.286	11	30	107	0.745	0.002	5	55
41	0.917	0.286	1	80	108	0.508	0.002	0	50
42	0.870	0.286	10	85	109	0.370	0.002	8	50
43	0.872	0.286	4	50	110	0.684	0.002	5	30
44	0.844	0.286	3	65	111	0.458	0.002	6	50
45	0.796	0.286	10	70	112	0.373	0.002	3	50
46	0.724	0.236	3	50	113	0.629	0.002	18	10
47	0.597	0.236	10	50	114	0.506	0.000	5	80
48	0.698	0.236	5	50	115	0.444	0.000	12	50
49	0.801	0.222	6	25	116	0.482	0.000	10	50

Continues

Appendix A: Continued

Subject	Consensus	Calibration	Years' experience	Self-proclamation	Subject	Consensus	Calibration	Years' experience	Self-proclamation
50	0.567	0.185	4	50	117	0.282	0.000	2	50
51	0.540	0.185	3	75	118	0.265	0.000	6	70
52	0.638	0.185	6	50	119	0.673	0.000	0.4	96
53	0.610	0.185	5	60	120	0.552	0.000	5	95
54	0.685	0.185	8	50	121	0.166	0.000	4	50
55	0.694	0.185	11	30	122	0.555	0.000	5	50
56	0.632	0.185	3	40	123	0.292	0.000	5	40
57	0.687	0.177	4	40	124	0.489	0.000	7	33
58	0.677	0.177	6	50	125	0.610	0.000	5	75
59	0.702	0.177	7	30	126	0.656	0.000	5	30
60	0.735	0.177	5	85	127	0.354	0.000	8	50
61	0.845	0.145	5	50	128	0.495	0.000	4	66
62	0.782	0.145	3	75	129	0.785	0.000	2	50
63	0.705	0.145	4	50	130	0.634	0.000	8	70
64	0.720	0.145	8	10	131	0.686	0.000	3	51
65	0.781	0.145	7	50	132	0.437	0.000	4	40
66	0.748	0.145	5	70	133	0.483	0.000	20	50
67	0.809	0.145	3	35	134	0.590	0.000	10	50
Subject	Consensus	Calibration	Years' experience	Self-proclamation					
135	0.335	0.000	3	60					
136	0.248	0.000	5	43					
137	0.336	0.000	3	40					
138	0.455	0.000	2	0					
139	0.272	0.000	6	85					
140	0.382	0.000	10	10					
141	0.612	0.000	4	50					
142	0.500	0.000	4	60					
143	0.401	0.000	6	75					
144	0.371	0.000	3	50					
145	0.488	0.000	8	50					
146	0.326	0.000	7	57					
147	0.526	0.000	4	50					
148	0.659	0.000	8	40					

149	0.557	0.000	5	50
150	−0.025	0.000	5	60
151	0.629	0.000	2	80
152	0.366	0.000	6	35
153	0.515	0.000	6	50
154	0.571	0.000	3	50
155	0.559	0.000	6	50
156	0.419	0.000	7	50
157	0.052	0.000	8	10
158	0.123	0.000	4	71
159	0.097	0.000	8	50
160	0.263	0.000	8	40
161	0.156	0.000	9	30
162	0.111	0.000	15	1
163	0.429	0.000	10	50
164	0.184	0.000	8	10

Appendix B: Measurements for Study B

Subject	Consensus	Calibration	Years' experience	Self-proclamation
1	0.544	0.132	6	40
2	0.708	0.385	5	50
3	0.666	0.132	4	10
4	0.707	0.370	9	75
5	0.750	0.385	4	30
6	0.737	0.132	9	15
7	0.539	0.083	2	50
8	0.516	0.025	3	95
9	0.541	0.018	6	50
10	0.558	0.018		50
11	0.657	0.049	5	27
12	0.495	0.049	3	50
13	0.578	0.011	3	60

(Continues)

Appendix B: Continued

Subject	Consensus	Calibration	Years' experience	Self-proclamation
14	0.359	0.078	6	85
15	0.383	0.000	2	60
16	0.446	0.001	3	10
17	0.519	0.000	10	10
18	0.338	0.000	6	40
19	0.195	0.000	10	30
20	0.640	0.080	5	50
21	0.298	0.000	3	50
22	0.472	0.011	10	75
23	0.402	0.000	3	50
24	0.420	0.007	11	3
25	0.108	0.000	10	20
26	0.337	0.000	5	50
27	0.159	0.000	7	66
28	0.129	0.000	8	30
29	0.485	0.083	10	15
30	0.541	0.031	8	20
31	0.321	0.000	5	50
32	0.357	0.001		50
33	0.250	0.000	3	25
34	0.447	0.002	5	70
35	0.443	0.000	5	50

Appendix C: Measurements for Study C

Subject	Consensus	Calibration	Years' experience	Self-proclamation
1	0.628	0.06362	4	60
2	0.614	0.01488	25	50
3	0.676	0.0005192	5	90
4	0.772	0.0007985	12	25
5	0.290	0.00001257	5	90
6	0.550	0.00002789	7	33
7	0.725	0.000001837	5	25
8	0.388	$3.211 * 10^{-14}$	1.5	50

9	0.432	$7.543 * 10^{-12}$	5	40
10	0.367	0.0000638	2	80
11	0.536	$8.734 * 10^{-9}$	5	90
12	0.470	$8.734 * 10^{-9}$	5	50
13	0.480	0.0003005	6	50
14	0.488	$3.376 * 10^{-8}$	5	50
15	0.516	0.0002767	8	35
16	0.647	0.00002789	5	10
17	0.637	0.0002767	4	90
18	0.588	0.00002789	5	74
19	0.618	$1.157 * 10^{-8}$	12	20

Appendix D: Measurements for Study D

Subject	Consensus	Calibration	Years' experience	Self-proclamation
1	0.611	0.615	5	70
2	0.675	0.615	10	70
3	0.736	0.615	10	99
4	0.755	0.492	15	5
5	0.462	0.385	10	75
6	0.386	0.370	10	25
7	0.365	0.313	7	50
8	0.221	0.313	10	30
9	0.437	0.197	5	50
10	0.481	0.197	6	50
11	0.645	0.197	5	33
12	0.394	0.197	5	80
13	0.156	0.154	5	50
14	0.683	0.068	8	80
15	0.646	0.068	3	50
16	0.651	0.001	10	80

Appendix E: Questions used to measure calibration

Survey Questions		
#	Questions in Study A	Realization (%)
1	If one of the seven NMAP commands was randomly selected and then executed, how probable do you think it is that a default configured Snort intrusion detection system would detect it?	72
2	If one of the seven NMAP commands was randomly selected and then executed, how probable do you think it is that a default configured Tamandua intrusion detection system would detect it?	29
3	If one of the seven NMAP commands was randomly selected and then executed, how probable do you think it is that a default configured Firestorm intrusion detection system would detect it?	29
4	Consider vulnerabilities of high severity (according to CVSS) that impacts Windows 7 and was published during 2010. What portion of these vulnerabilities has a corresponding signature in Snort's default ruleset?	40
5	Consider vulnerabilities of high severity (according to CVSS) that impacts MySQL and was published during 2004–2009. What portion of these vulnerabilities has a corresponding signature in Snort's default ruleset?	87
6	Consider vulnerabilities of high severity (according to CVSS) that impacts Windows 7 and was published during 2009. What portion of these vulnerabilities has a corresponding signature in Snort's default ruleset?	37
7	Consider vulnerabilities of high severity (according to CVSS) that impacts Windows 7 and was published during the last 6 months. What portion of these vulnerabilities has a corresponding signature in Snort's default ruleset?	35
8	Consider vulnerabilities of high severity (according to CVSS) that impacts Samba and was published during 2010. What portion of these vulnerabilities has a corresponding signature in Snort's default ruleset?	33
#	Questions in Study B	Realization (%)
1	What is the share of known vulnerabilities with some impact on availability?	71
2	Of the known vulnerabilities with some impact on availability, how large portion can be exploited from external networks?	85
3	Of the known vulnerabilities with some impact on availability, how large portion requires that the attacker can bypass authentication?	5
4	What is the share of known vulnerabilities with some impact on availability that affect Windows 7?	85
5	What is the share of known vulnerabilities with complete impact on availability?	23
6	What portion of organizations in EMEA and US that operate their business online has an important online reputation use some on-premise/in-house DDoS protection technology?	65

7	What portion of organizations in EMEA and US that operate their business online or have an important online reputation over provision their bandwidth to protect against potential DDoS threats?	28
8	What portion of organizations in EMEA and US that operate their business online, have an important online reputation or operate financial services are primarily suffering from target DDoS attacks and aware of whom the attackers are?	30
9	What portion of organizations in EMEA and US that operate their business online or have an important online reputation or operate online financial services is primarily suffering from random DDoS?	52
10	What portion of organizations in EMEA and US that operate their business online or have an important online have experienced a DDoS attacks during a year that did disrupt services?	31
11	What portion of organizations in EMEA and US that operate their business online, has an important online have experienced and has experienced DDoS attacks needed more than 5 hours to recover from the most severe attack?	41

#	Questions in Study C	Realization (%)
1	How many of the high-severity vulnerabilities published in 2010 have a full impact on Confidentiality, Integrity and Availability?	57
2	How many of the medium-severity vulnerabilities published in 2010 have a full impact on Confidentiality, Integrity and Availability?	6
3	How many of the vulnerabilities published in 2010 that can be exploited remotely require that the attacker bypass some authentication mechanism first?	9
4	How many of the vulnerabilities published in 2010 that can be exploited remotely and require that the attacker bypass some authentication mechanism first is of severity-rating high?	15
5	How many of the vulnerabilities published in 2010 that can be exploited remotely are of severity-rating high?	52
6	What is the probability that an attack (selected randomly from the 20 listed) will be prevented if Libverify and Libsafe are used?	0
7	What is the probability that an attack (selected randomly from the 20 listed) will be halted if Libverify and Libsafe are used?	20
8	What is the probability that an attack (selected randomly from the 20 listed) will be prevented if ProPolice is used?	40
9	What is the probability that an attack (selected randomly from the 20 listed) will be halted if ProPolice is used?	10

(Continues)

Appendix E: Continued

Survey Questions		
10	What is the probability that an attack (selected randomly from the 20 listed) will be prevented if Stackguard's terminator canary is used?	0
11	What is the probability that an attack (selected randomly from the 20 listed) will be halted if Stackguard's terminator canary is used?	15
#	Questions in Study D	Realization (%)
1	What portion of vulnerabilities published during 2010 of high severity has a complete impact on Confidentiality, Integrity and Availability?	57
2	What portion of vulnerabilities published during 2010 of medium severity has a complete impact on Confidentiality, Integrity and Availability?	6
3	What portion of vulnerabilities published during 2010 that are remotely exploitable (does not require LAN access) will require that the attacker can authenticate itself before succeeding with an exploit?	9
4	What portion of vulnerabilities published in 2010 that are remotely exploitable (does not require LAN access) and requires that the attacker can authenticate itself before the exploit is of high severity?	15
5	What portion of vulnerabilities published in 2010 that are remotely exploitable (does not require LAN access) is of high severity?	52
6	What portion of vulnerabilities publicly announced in 2010 with high severity is due to input validation or buffer errors?	53
7	What portion of vulnerabilities publicly announced with high severity for Windows 7 is due to input validation or buffer errors?	36
8	What portion of vulnerabilities publicly announced with high severity for Apple's products is due to input validation or buffer errors?	31
9	What portion of vulnerabilities publicly announced with high severity for the .NET framework is due to authentication or authorization errors?	10
10	What portion of vulnerabilities publicly announced with high severity for the Microsoft's Internet Information Services is due to authentication or authorization errors?	13
11	What portion of vulnerabilities publicly announced with high severity for Cisco's products is due to authentication or authorization errors?	11

References

ABDOLMOHAMMADI, M. J. and J. SHANTEAU (1992) Personal attributes of expert auditors, *Organizational Behavior and Human Decision Processes*, **53**, 158–172.

ALHAZMI, O. H., Y. K. MALAIYA and I. RAY (2007) Measuring, analyzing and predicting security vulnerabilities in software systems, *Computers & Security*, **26**, 219–228.

ASHTON, A. H. (1985) Does consensus imply accuracy in accounting studies of decision making? *The Accounting Review*, **60**, 173–185.

AXELSSON, S. (2000) Intrusion detection systems: a survey and taxonomy.

AYYUB, B. M. (2001) Elicitation of Expert Opinions for Uncertainty and Risks, Boca Raton, Florida, USA: CRC.

BISHOP, M. (2003) What is computer security? *Security & Privacy, IEEE*, **1**, 67–69.

BOLGER, F. and G. WRIGHT (1994) Assessing the quality of expert judgment: issues and analysis, *Decision Support Systems*, **11**, 1–24.

CAMPBELL, K., et al. (2003) The economic cost of publicly announced information security breaches: empirical evidence from the stock market, *Journal of Computer Security*, **11**, 431–448.

CAVUSGIL, S. T. and L. A. ELVEY-KIRK (1998) Mail survey response behavior: a conceptualization of motivating factors and an empirical study, *European Journal of Marketing*, **32**, 1165–1192.

CAVUSOGLU, H., B. MISHRA and S. RAGHUNATHAN (2004) The effect of internet security breach announcements on market value: capital market reactions for breached firms and internet security developers, *International Journal of Electronic Commerce*, **9**, 70–104.

CHASE, W. G. and K. A. ERICSSON (1981) Skilled memory, *Cognitive Skills and Their Acquisition*, 141–189.

CHI, M. (1988) The nature of expertise.

COOKE, R. (2008) TU Delft expert judgment data base, *Reliability Engineering and System Safety*, **93**, 657–674.

COOKE, R. (1991) Experts in Uncertainty: Opinion and Subjective Probability in Science, Oxford University Press: USA.

CRONBACH, L. J. (1951) Coefficient alpha and the internal structure of tests, *Psychometrika*, **16**, 297–334.

CRONBACH, L. J. and R. J. SHAVELSON (2004) My current thoughts on coefficient alpha and successor procedures, *Educational and Psychological Measurement*, **64**, 391–418.

DEN BRABER, F., et al. (2007) Model-based security analysis in seven steps – a guided tour to the CORAS method, *BT Technology Journal*, **25**, 101–117.

DODGE, R. C. (2007) Phishing for user security awareness, *Computers & Security*, **26**, 73–80.

EINHORN, H. J. (1974) Expert judgment: some necessary conditions and an example, *Journal of Applied Psychology*, **59**, 562.

EINHORN, H. J. (1972) Expert measurement and mechanical combination, *Organizational Behavior and Human Performance*, **7**(1), 86–106.

ELSEVIER, B. V. (2010) Scopus. Available at: http://www.scopus.com/ [Accessed September 30, 2010].

Elsevier Inc. (2010) Inspec, Compendex. Available at: http://www.engineeringvillage2.org/controller/servlet/Controller?CID=quickSearch&database=3 [Accessed September 30, 2010].

FARRINGTON-DARBY, T. and J. R. WILSON (2006) The nature of expertise: a review, *Applied Ergonomics*, **37**, 17–32.

FISCHER, G. (1981) When oracles fail – a comparison of four procedures for aggregating subjective probability forecasts, *Organizational Behavior and Human Performance*, **28**(1), 96–110.

GABLE, G. G. (1994) Integrating case study and survey research methods: an example in information systems, *European Journal of Information Systems*, **3**, 112–126.

GARTHWAITE, P. H., J. B. KADANE and A. O'HAGAN (2005) Statistical methods for eliciting probability distributions, *Journal of the American Statistical Association*, **100**, 680–701.

GIGERENZER, G. and P. M. TODD (1999) Simple Heuristics that Make us Smart, USA: Oxford University Press.

GLASS, G. V. and J. C. STANLEY (1970) Statistical Methods in Education and Psychology. Englewood Cliffs. N. Y.: Prentice-Hall.

GOLDBERG, L. R. (1968) Simple models or simple processes? Some research on clinical judgments, *American Psychologist*, **23**, 483.

GROVES, R. M., et al. (2009) Survey Methodology, Hoboken, New Jersey, USA: John Wiley & Sons Inc.

HAIMES, Y. Y. (2003) Accident precursors, terrorist attacks, and systems engineering. In *NAE Workshop*.

HANSMAN, S. and R. HUNT (2005) A taxonomy of network and computer attacks, *Computers & Security*, **24**, 31–43.

HOENIG, M. (1985) Drawing the Line on Expert Opinions, *Journal of Products Liability* **8**, 335–336.

HOLM, H., et al. (2011) Expert assessment on the probability of successful remote code execution attacks. In *The 8th International Workshop on Security in Information Systems*.

HUMPHREYS, T. (2006) State-of-the-art information security management systems with ISO/IEC 27001: 2005. *ISO Management Systems*, **6**, 15–18.

JANES, J. (1999) Survey construction, *Library hi tech*, **17**, 321–325.

LAPRIE, J. C., B. RANDELL and C. LANDWEHR (2004) Basic concepts and taxonomy of dependable and secure computing, *IEEE Transactions on Dependable and Secure Computing*, **1**, 11–33.

MADAN, B. B., et al. (2002) Modeling and quantification of security attributes of software systems. in International Conference on Dependable Systems and Networks. IEEE Computer Society.

MCHUGH, J. (2000) Testing Intrusion detection systems: a critique of the 1998 and 1999 DARPA intrusion detection system evaluations as performed by Lincoln Laboratory, *ACM Transactions on Information and System Security*, **3**, 262–294.

PHELPS, R. H. (1978) Expert Livestock Judgment: A Descriptive Analysis of the Development of Expertise, Kansas State University, Manhattan, Kansas, USA: ProQuest Information & Learning.

SHANTEAU, J. (1988) Psychological characteristics and strategies of expert decision makers, *Acta Psychologica*, **68**, 203–215.

SHANTEAU, J. (2001) What does it mean when experts disagree, *Linking expertise and naturalistic decision making*, 229–244.

SHANTEAU, J., et al. (2002) Performance-based assessment of expertise: how to decide if someone is an expert or not, *European Journal of Operational Research*, **136**, 253–263.

SOMMESTAD, T., M. EKSTEDT and P. JOHNSON (2010) A probabilistic relational model for security risk analysis, *Computers & Security*, **29**, 659–679.

SOMMESTAD, T., H. HOLM and M. EKSTEDT (2011) Estimates of success rates of Denial-of-Service attacks. 2011 IEEE 10th International Conference on Trust, Security and Privacy in Computing and Communications (TrustCom).

SOMMESTAD, T., H. HOLM and M. EKSTEDT (2012a) Estimates of success rates of remote arbitrary code execution attacks, *Information Management & Computer Security*, **20**, 107–122.

SOMMESTAD, T., H. HOLM and M. EKSTEDT (2012b) Effort estimates for vulnerability discovery projects. 2012 45th Hawaii International Conference on System Science (HICSS).

SOMMESTAD, T., H. HOLM, M. EKSTEDT and N. HONETH (Submitted) Quantifying the effectiveness of intrusion detection systems in operation through domain experts.

TAYLOR, C., A. KRINGS and J. ALVES-FOSS (2002) Risk analysis and probabilistic survivability assessment (RAPSA): an assessment approach for power substation hardening. In Proc. ACM Workshop on Scientific Aspects of Cyber Terrorism, (SACT), Washington DC.: Citeseer.

TRUMBO, D., et al. (1962) Reliability and accuracy in the inspection of hard red winter wheat, *Cereal Science Today*, **7**, 62–71.

VERENDEL, V. (2009) Quantified security is a weak hypothesis: a critical survey of results and assumptions, *Proceedings of the 2009 workshop on New security paradigms workshop*.

WARNER, R. M. (2008) Applied Statistics: From Bivariate Through Multivariate Techniques, Thousand Oaks, California, USA: Sage Publications, Inc.

WEISS, D. and J. SHANTEAU (2003) Empirical assessment of expertise, *Human Factors: The Journal of the Human Factors and Ergonomics Society*, **45**, 104–116.

The authors

Hannes Holm

Hannes Holm is a PhD student at the department of Industrial Information and Control Systems at the Royal Institute of Technology (KTH) in Stockholm, Sweden. He received his MSc degree in management engineering at Luleå University of Technology. His research interests include enterprise security architecture and cyber security regarding critical infrastructure control systems.

Teodor Sommestad

Teodor Sommestad received his PhD degree in 2012 Industrial Information and Control Systems his MSc degree in Computer Science in 2005, both at the Royal Institute of Technology (KTH) in Stockholm, Sweden. He is currently a senior scientist at the Swedish Defence Research Agency (FOI), Linköping, Sweden. Teodor is involved in a number of projects addressing decision-making problems related to cyber security. Among other thing, Teodor is involved in research on information security culture and research related to cyber ranges and cyber security exercises.

Mathias Ekstedt

Mathias Ekstedt is an associate professor at the Royal Institute of Technology (KTH) in Stockholm, Sweden. His research interests include systems and enterprise architecture modelling and analyses with respect to information and cyber security, in particular for the domain of Power system management. He is the manager of the programme IT Applications in Power System Operation and Control within the Swedish Centre of Excellence in Electric Power Engineering and technical coordinator of the EU FP7 project VIKING. He received his MSc, PhD and Docent from the Royal Institute of Technology in 1999, 2004 and 2010, respectively.

Nicholas Honeth

Nicholas Honeth, received the MSc degree in computer science from Chalmers University of Technology, Gothenburg, Sweden, and the BSc degree in electrical and computer engineering from the University of Cape Town, Republic of South Africa. He is currently a PhD student at the department of Industrial Information and Control Systems at KTH – The Royal institute of Technology, Stockholm, Sweden. His chief interests are in intelligent control systems for electrical distribution networks.