The Significance of Information Security Risk Assessments

Exploring the Consensus of Raters' Perceptions of Probability and Severity

Jonas Hallberg, Johan Bengtsson, Niklas Hallberg, Henrik Karlzén, Teodor Sommestad
Swedish Defence Research Agency, P.O. Box 1165, SE-581 11 Linköping, Sweden

Abstract - *Identifying and assessing risks is vital in striving for adequate information security. The basis for the assessments is the probability and the severity of possible incidents affecting the confidentiality, integrity, and availability of information assets. However, assessing the probability and the severity of possible events is not straightforward. The objective of this paper is to explore the consensus of raters assessing the probability and the severity of information security incidents. Data collected through questionnaires are used to evaluate the consensus of 20 raters when assessing 105 information security incidents. The results indicate that the consensus of the raters is too low for the assessment results to provide a sound basis for decisions. In conclusion, better support is needed for assessing information security risks in order to reach the necessary consensus levels.*

Keywords: Information security, Risk assessment, Inter-rater reliability, Consensus

1 Introduction

Organizations' management of information security should be based on the information security risks they are facing [1]. However, to assess the information security risks is challenging, due to the dynamic nature of the information assets and systems as well as the organizations' security efforts. In these kinds of situations, where there are no straightforward answers readily available, the opinions, or perceptions, of experts are needed. Unfortunately, even high consensus among the experts does not guarantee that their answers are correct; they may be uniformly wrong in their assessments. Thus, the consensus of experts is required, but not a guarantee, for valid results [2]–[5] and has been studied in several different areas, such as, medical pathology [2], [4], accounting [4], [5], psychology [4], [6], [7], and cyber security [3].

Despite the substantial number of methods and frameworks proposed for assessing information security risks based on perceived probabilities and severities, there are no methods available alleviating the issue of insufficient reliability (and consequently also lacking validity) in the assessments [8]. The challenges faced when striving for reliable assessments are highlighted by the vast range of input proposed for assessment methods and frameworks [9].

The objective of this paper is to explore the consensus (inter-rater reliability) of raters' individual assessments of the probability and the severity of information security incidents. This is motivated by the difficulty to directly assess the validity of the results and the requirement of reliability for validity [2]–[5]. The addressed research question is: Are raters' assessments of the probability and the severity of information security incidents reliable enough to be used as a basis for information security management? In order to answer the research question, the following two hypotheses are tested.

H1. According to proposed measures of consensus, ratings of the probability of information security incidents are reliable enough to serve as the basis for decisions.
H2. According to proposed measures of consensus, ratings of the severity of information security incidents are reliable enough to serve as the basis for decisions.

If the results do not show perfect consensus, there may be several factors causing the lack of agreement. Two variables that may affect the results are whether the raters are experts in the areas of information security and risk assessment and the cognitive style of the raters. Thus, the following two hypotheses are tested.

H3. Information security and risk assessment experts reach higher consensus than non-experts.
H4. Raters with a logical cognitive style reach higher consensus than raters with an intuitive cognitive style.

There may also be other factors that make some raters especially bad at assessing the probability and the severity of information security incidents. Alternatively, the specification of some of the incidents might render them difficult to assess and, thereby, substantially lower the overall consensus of the raters (when a set of incidents is considered). Thus, the following two hypotheses are tested.

H5. Removing a fraction of the raters yields a subset of raters whose consensus is substantially higher.
H6. Removing a fraction of the incidents yields a subset of incidents for which the consensus of the raters is substantially higher.

The rest of the paper is structured as follows. Section 2 and 3 describe the method and results respectively. Section 4 discusses the results and section 5 concludes the paper.

2 Method

In this study, consensus measures are applied to data on the assessment of the probability and the severity of information security incidents. The assessment data were collected in two previous studies [10], [11]. The subsections below describe the participants included and the questionnaire used in those studies as well as the computation of the consensus measures.

2.1 Participants and questionnaire

The questionnaire was distributed to a strategic sample of 20 researchers in the areas of IT security, IT management, and human factors at the Swedish Defence Research Agency (the authors' own organization). All the respondents possess university degrees, are in the age range 29 to 64 years, work as researchers, and are familiar with the concepts of probability and severity assessments. In this study, the expertise, considering information security and risk assessment, is used as a variable to analyze whether it affects the results. Thus, the participants being merely familiar with the concepts of probability and severity assessments can be distinguished from the experts, or at least the self-proclaimed experts. In addition to expertise, the questionnaire includes items used to measure the cognitive style. Based on their answers, the participants are graded on a scale where the two ends represent logical and intuitive cognitive style respectively.

The questionnaire includes 105 incidents. For each incident, the probability and the severity were marked on visual analog scales ranging from 0 to 100% and 0 to 10 respectively. Considering probability, the value represents the probability of the incident occurring within the following ten years. Considering severity, the value 0 represents minimal or no harm at all, whereas 10 represents the greatest harm caused by any of the incidents. It was also specified that the scale is proportional, i.e., 5 is half as harmful as 10. For efficiency and precision, the answers were measured in millimeters from the anchor marking the value 0, instead of the values used on the scales. This resulted in values ranging from 0 to 108, where 108 corresponds to the probability value 100% and the severity value 10, respectively.

The incidents were designed to be meaningful for the target population. For example, they used information assets and incidents that are relevant for the organization. An example of the incidents is: *A scientist's USB-stick with five years of collected (unclassified) material is stolen at an international conference.*

The outcome of the performed assessments is illustrated in Fig. 1, where the results of the 20 respondents' assessment of the 105 incidents (in total yielding 2,100 incident assessments, each consisting of one probability and one severity value) are plotted after being mapped into scales from 1 to 10. The graph illustrates that the distribution of the assessments is far from uniform, although there are values spread over the whole sample space.

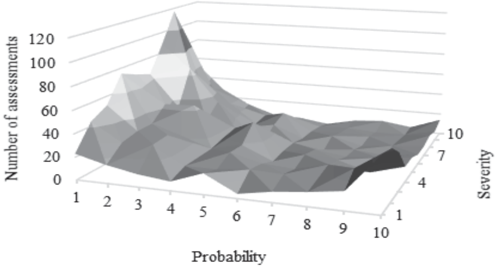

Fig. 1. The distribution of the pairs of probability and severity values.

2.2 Computation of consensus measures

There are several methods for measuring the consensus of raters, also referred to as inter-rater reliability indices [12]. A prominent method is Krippendorff's alpha (KA) that handles data measured on any of several different kinds of scales (e.g., ordinal, interval, or ratio [13]), an arbitrary number of raters (larger than 1), and missing data [14]. KA values are computed according to the equation:

$$KA = 1 - (D_d/D_a) \tag{1}$$

where D_d is *the average disagreement between raters regarding the units rated.* A difference function, which has to be adapted to the type of scale used during the rating [14], is used to compute the average difference between all the values assigned to each unit (incident). D_d is computed as the average of the averages computed for each unit. D_a is *the average difference between all the values assigned by the raters* and computed using the same difference function as for computing the disagreement.

The values returned by the difference function are 0 or larger. Thus, the maximal KA value equals 1 and represents perfect agreement between the raters. The KA value decreases with increasing disagreement between the raters and the KA value 0 represents random rating. In extreme cases, with systematic disagreements, negative KA values may appear. Considering not only the disagreement but also the average difference provides two major advantages; the index is adjusted in situations where low disagreement depends on skewed use of the rating values, and the effect of lower disagreement resulting from using scales with fewer steps is removed. Since KA requires absolute agreement, i.e., does not compensate for systematic bias, it is occasionally being referred to as a conservative index [15].

In this study, the difference function for interval scales was used, i.e., the difference between values was squared during the computation of the average disagreements and the average differences. To automate the many computations of KA values, a script was implemented in Microsoft Excel (2013). The script was validated by the use of test cases and comparison to the output from the online tool ReCal OIR [16] and the SPSS macro KALPHA [17].

The results provided by KA are complemented with the Pearson product-moment correlation coefficient [13], hereafter referred to as the Pearson coefficient (PC). The PC is used to acquire measures of consensus based on the covariation between raters and, consequently, does not penalize for systematic bias. Assuming that calibration of raters can remove systematic bias, the relative weighing of the incidents becomes a central aspect when targeting consensus. Still, the use of covariation measures when assessing rater consensus is not considered appropriate by all researchers [15]. Despite the criticism, the PC is used because of the possibility of calibration and to be able to compare the results to studies performed in other areas.

The consensus measures based on the PC were computed in two steps. First, the PCs for all pairs of raters were computed. Second, the consensus measure was computed as the average of all the computed PCs. In order to support the large number of combinations arising from dividing the set of raters into different subsets, a script was implemented in Microsoft Excel (2013).

3 Results

The results include the data on consensus used to test the formulated hypotheses (H1 to H6).

3.1 Overall inter-rater reliability

The KA and PC values were computed from the 2,100 assessments of the probability and the severity of information security incidents. Table 1 includes the results.

Table 1. KA and PC values

Consensus coefficient	Probability	Severity
KA	0.30	0.42
PC	0.42	0.54

For consensus coefficients to be useful, reference values are needed to compare the results with. These reference values may be based on experience or the results of others. For the KA, Krippendorff [18] introduced the interpretation that results in the interval 0.8 to 1 indicate reliable data and results in the interval 0.667 to 0.8 indicate that the data should be used with care, whereas all data resulting in KA values below 0.667 should be discarded. For the PC, the results presented in some previous studies on the correlation between experts in different fields of expertise are used as reference, Table 2.

Table 2. Previously reported consensus coefficients

Expert	Result	Study
Weather forecaster	0.95	[4]
Auditor	0.76	[4]
Violence risk assessment	0.76	[6]
Cyber security, intrusion detection systems	0.64	[3]
Cyber security, arbitrary code execution attacks	0.56	[3]
Pathologist	0.55	[4]
Cyber security, software vulnerability discovery	0.54	[3]
Cyber security, denial of service attacks	0.48	[3]
Clinical psychologist	0.40	[4]
Stockbroker	0.32	[4]

Consequently, based on the KA values, the data should *not* be used for any decisions, that is, the hypotheses H1 and H2 are refuted. Based on the PCs, the consensus of the raters is at the same levels as for experts in clinical psychology and software vulnerability discovery considering probability and severity respectively. Thus, there is no clear answer to whether the hypotheses H1 and H2 are supported or not.

3.2 Expertise and cognitive style

Two variables that may affect the consensus coefficients are the level of expertise and the cognitive style. Table 3 includes the KA values for the different subsets of raters classified as: information security and risk assessment experts and non-experts respectively, as well as logic and intuitive cognitive style respectively.

Table 3. The KA and PC values for different subsets of raters based on expertise and cognitive style

Rater subset	Raters in set	Probability	Severity
Experts	10	0.25	0.44
Non-experts	10	0.37	0.41
Logical	9	0.27	0.27
Intuitive	11	0.29	0.55

No PCs were calculated for the subsets of raters called experts, non-experts, logical, and intuitive. Instead regression analysis was performed to identify significant regression models considering the answers to the questions on reasoning and expertise, on one hand, and the average PC between each rater and all the other raters. However, no significant relations were found.

Based on the KA values in Table 3, the consensus of experts is clearly lower considering probability and slightly higher considering severity than the consensus of non-experts. Thus, the hypothesis H3 is not supported. The consensus of raters with logic cognitive style is slightly lower considering probability and drastically lower considering severity than the consensus of raters with intuitive cognitive style. Thus, the hypothesis H4 is not supported. Since no significant relations were found, neither the regression models support the hypotheses H3 and H4.

3.3 Removing incidents and raters to improve consensus

This section includes results supporting the analysis of whether removing incidents and raters yields substantial improvements in the consensus. The results include illustrations of the variation of the computed disagreement and PC values, a method for the adjustment of the KA values yielded by removing incidents and raters, and the KA and PC values based on the removal of incidents and raters.

3.3.1 Disagreement values and Pearson coefficients

The disagreement values, D_d in equation (1), computed for each of the incidents vary widely. Fig. 2 illustrates the values provided by the 20 raters for the incidents with the highest (incident 100) and lowest (incident 40) disagreement values for probability. Further illustration of the distribution of the disagreement values for the incidents is provided in Fig. 3, where ten bins are used to illustrate the distribution for severity. The number of incidents in a bin corresponds to the number of incidents whose disagreement value is larger than the bin maximum value of the previous bin and smaller than or equal to the bin maximum value of the current bin, e.g., all incidents placed in the second bin have a disagreement value over 290 and below or equal to 580.

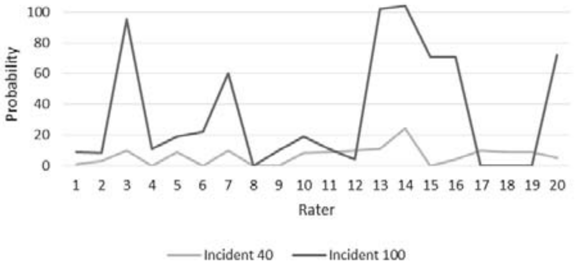

Fig. 2. The incidents with the lowest and the highest disagreements for probability. Because of the measurement method, the largest value is 104.

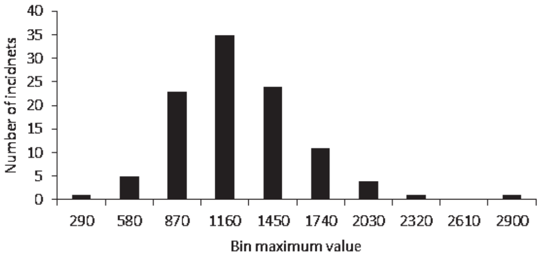

Fig. 3. The distribution of disagreement values for severity.

The variation in disagreement values between the incidents supports the general assumption leading to hypothesis H5. That is, removing selected incidents will increase the consensus for the remaining incidents.

Fig. 4, illustrates the distribution of each raters' average PC with the other raters. The variation in the average PC supports the general assumption leading to hypothesis H6. That is, removing selected raters will increase the consensus for the remaining raters.

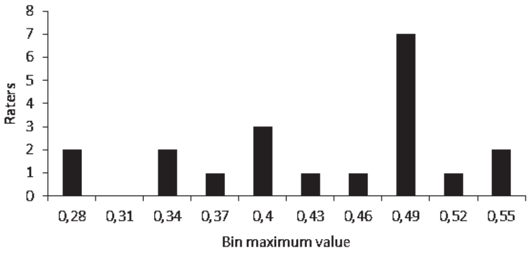

Fig. 4. The distribution of raters' average PC.

3.3.2 Adjusting the KA values

A viable question in relation to the improvement of KA values by removing incidents and raters is to what extent the improved values are due to the incidents and raters being removed with hindsight. To address the issues the increase of the KA values is assumed to stem from two factors: (1) the higher consensus between the remaining raters for the remaining incidents and (2) the possibility to retain the values that in retrospect provide the best KA values. In this study, the second factor is referred to as the hindsight effect.

To model the improvement of the KA values, a factor is introduced in equation (1). Thus, the following equation is proposed to specify the improved KA.

$$KA' = 1 - (1 - I) * C \qquad (2)$$

where *KA'* is the improved KA, *I* is the improvement, and *C*, which is introduced to simplify the equations, is the fraction between the average disagreements (D_d) and the average differences (D_a). If the improvement equals zero then *KA'* will equal the original KA. If the improvement equals 1 then also *KA'* will equal 1, i.e. the maximal KA value. Thus, assuming that the improvement is not negative, the improvement values (*I*) are in the interval 0 to 1. Assuming that the two causes for the improvement overlap, the following equation is used to model their relationship.

$$I = I_c + I_h - I_c * I_h \qquad (3)$$

where *I* is the improvement of the KA values yielded by removing incidents and raters, I_c is the improvement caused by the actual improvement of the consensus among the remaining raters for the remaining incidents, and I_h is the improvement caused by hindsight effects. Reformulating equation (3) yields:

$$I_c = (I - I_h) / (1 - I_h) \qquad (4)$$

Using equation (2), the adjusted KA quantifying the actual improvement in the consensus between raters is described by the following equation.

$$KA_c' = 1 - (1 - I_c) * C \qquad (5)$$

It may be noted that for the original KA, when no incidents or raters have been removed, there is no hindsight effect so KA_h should be zero and KA_c is equal to *KA*. Combining equation (4) and (5) results in the equation:

$$KA_c' = 1 - C * (1 - I) / (1 - I_h) \qquad (6)$$

Reformulating equation (2) to $C * (1 - I) = 1 - KA'$ yields the equation:

$$KA_c' = 1 - (1 - KA') / (1 - I_h) \qquad (7)$$

Using equation (2) for the improvement related to the hindsight effect results in $(1 - I_h) = (1 - KA_h') / C_h$, where KA_h' is the improved KA value achieved by removing values from a set of random assessment values, and C_h corresponds to the *C* value for random assessment values (and should be equal to 1 for truly random values). Using this and $C_h = 1 - KA_h$ yields the equation:

$$KA_c' = 1 - (1 - KA_h) * (1 - KA') / (1 - KA_h') \qquad (8)$$

Thus, KA_c can be computed from the KA values computed for the collected data, yielding KA', and random data, yielding KA_h and KA_h'.

3.3.3 KA when incidents and raters are removed

Table 4 contains the KA values for the probability and the severity assessments respectively. Each value corresponds to the KA for a combination of a set of raters and a set of incidents. The first row of results in the table is based on all the incidents. The following rows include the results for five subsets resulting from removing, one by one, the incidents causing the largest reduction in the KA values until the specified fraction of incidents remains. Correspondingly, the first column of the results in Table 4 contains the results considering all the raters. The following columns include the results for three subsets resulting from removing, one by one, the raters causing the largest reduction in the KA values until the specified fraction of raters remains. For the sets where both incidents and raters have been removed, the incidents were removed first.

To be able to compute adjusted KA values according to equation (8), the KA values achieved when removing incidents and raters from sets of random assessment values are needed. Table 5 includes the average KA values yielded by removing incidents and raters from 1,000 sets of random assessment values. Ideally, the KA value when random values are assigned for all incidents and raters (top left in Table 5) should be 0, whereas the computed average is -0.0003. The 95% confidence interval for the KA value is $-7.6 \cdot 10^{-4}$ to $9.6 \cdot 10^{-5}$ and thus includes 0. As illustrated by the data, the KA will increase when incidents and raters are removed, even if the values assigned by the raters are random. This increase is due to the hindsight effect discussed in the previous subsection.

Table 4. KA values for probability and severity

Inci-dents	Probability KA values				Severity KA values			
	All raters	75%	50%	25%	All raters	75%	50%	25%
All	0.30	0.42	0.52	0.58	0.42	0.54	0.63	0.73
90%	0.33	0.46	0.55	0.59	0.46	0.58	0.68	0.78
75%	0.39	0.52	0.60	0.65	0.52	0.64	0.74	0.83
50%	0.51	0.64	0.72	0.76	0.62	0.73	0.81	0.86
25%	0.71	0.78	0.83	0.89	0.72	0.85	0.90	0.92
10%	0.86	0.92	0.97	0.99	0.81	0.91	0.95	0.97

Table 5. Average KA values for 1,000 sets of random data

Incidents	Raters			
	All	75%	50%	25%
All	-0.0003	0.0220	0.0519	0.1115
90%	0.0063	0.0302	0.0622	0.1257
75%	0.0181	0.0465	0.0833	0.1533
50%	0.0454	0.0846	0.1323	0.2175
25%	0.0985	0.1637	0.2377	0.3565
10%	0.1642	0.2733	0.3897	0.5524

Table 6 includes the adjusted KA values provided by equation (8) when applied to the KA values in Table 4 and 5. The adjusted values are fairly close to the original values in Table 4, for probability the maximal decrease is below 10% and for severity the maximal decrease is below 5%. This is explained by the limited size of the KA values for random data when few incidents and raters have been removed as well as the values in Table 5 being relatively low compared to the values in Table 4 also when large fractions of incidents and raters are removed.

Like KA, the PCs were computed for several different sets of incidents and raters (Table 7). For each combination of the sets of raters and incidents, the PCs were computed for each pair of raters. The reliability measure was then computed as the average of the values for each pair of raters.

The result in Table 6 shows that removing limited fractions of the incidents or raters (maximum 25%) will not, according to Krippendorff's interpretation [18], yield acceptable KA. Thus, the hypotheses H5 and H6 are not supported. Considering the PCs in Table 7, there is an increase to at most 0.53 for probability and 0.64 for severity (both results are achieved by removing 25% of the incidents). Overall, the PCs provides some support for the hypotheses as the results correspond to the consensus of experts in software vulnerability discovery and intrusion detection systems respectively (Table 2).

Table 6. Adjusted KA values (KA_C) for probability and severity

Inci-dents	Probability KA values				Severity KA values			
	All raters	75%	50%	25%	All raters	75%	50%	25%
All	0.30	0.41	0.50	0.53	0.42	0.53	0.61	0.69
90%	0.33	0.44	0.52	0.54	0.46	0.57	0.66	0.74
75%	0.38	0.50	0.56	0.58	0.51	0.62	0.72	0.80
50%	0.49	0.60	0.68	0.69	0.60	0.70	0.78	0.82
25%	0.67	0.73	0.78	0.83	0.69	0.82	0.86	0.88
10%	0.83	0.89	0.94	0.97	0.77	0.87	0.92	0.94

Table 7. The PCs for probability and severity

Inci-dents	Probability PCs				Severity PCs			
	All raters	75%	50%	25%	All raters	75%	50%	25%
All	0.42	0.51	0.58	0.65	0.54	0.60	0.67	0.74
90%	0.47	0.55	0.63	0.70	0.59	0.65	0.71	0.79
75%	0.53	0.60	0.68	0.75	0.64	0.71	0.77	0.84
50%	0.65	0.72	0.77	0.83	0.72	0.78	0.83	0.88
25%	0.79	0.85	0.89	0.92	0.81	0.87	0.91	0.94
10%	0.92	0.97	0.98	0.99	0.89	0.95	0.97	0.98

4 Discussion

In this section, the results presented in the previous section are discussed, considering the overall consensus, the influence of expertise and cognitive style, the removal of incidents and raters, and the limitations of the study.

4.1 Overall inter-rater reliability

Based on the KA values achieved and the interpretation of KA values introduced by Krippendorff [18], no information security management decisions should be based on the assessment data. Considering the PCs, the raters are performing like experts in clinical psychology, when assessing probability, and like experts in software vulnerability discovery, when assessing severity (Table 7). Thus, according to the KA values, the hypotheses H1 and H2 are not supported, whereas the situation is unclear when the PCs are considered. Still, the PCs indicate that the ratings of probability and severity are not reliable enough between raters to be considered a sound basis for the quantification of information security risks.

4.2 Expertise and congnitive style

The data used in this study include assessments by experts as well as non-experts. The KA values computed for these two sets respectively (Table 3) indicate small differences considering severity but considerable differences considering probability (about 50%), implying that the non-experts have a higher consensus than the experts. However, considering the regression model for the average PC between each rater and all the other raters, and the answers to the questions on expertise, there is no significant relation between the constructs. Consequently, neither the KA values nor the PCs show that the experts have a higher consensus than the non-experts and the third hypothesis (H3) is not supported. The implication is that it cannot be stated that experts have a higher consensus than non-experts when the probability and the severity of information security incidents are rated.

Considering cognitive style, the raters have been divided into two subsets based on their answers to the related questions. The raters ending up in the subset with the most intuitive raters score a considerably higher KA than the raters in the other subset for the severity assessments. Still, the regression model for cognitive style and the PC coupled to the rating of severity did not yield a significant relation between the two constructs. Thus, the fourth hypothesis (H4) is not supported. The implication is that there is no indication of raters being more inclined to logical reasoning reaching higher consensus but rather the opposite considering severity, although no significant relation was found.

4.3 Removal of incidents and raters

The disagreement values for the incidents vary (as illustrated by Fig. 2 and Fig. 3). This supports the formulation of the hypotheses that some incidents are harder to assess than others (H6). Along these lines, the KA values yielded by removing incidents and raters one by one in order to increase the KA, show that the selected subsets of incidents and raters certainly provides higher values. However, there are issues related to this procedure. First, it could be argued that the increase in KA values is due to the removal of incidents and raters in retrospect, which in this study is referred to as the hindsight effect, rather than the removed incidents being hard to assess and the removed raters being poor assessors. To take the possibility of the hindsight effect into account, a set of adjusted KA values has been computed (Table 6). These values indicate that the KA values are increasing for the selected subsets even though the hindsight effect has been considered. Second, the rate at which the KA can increase when incidents are removed is limited. As illustrated in Fig. 3, the distribution of the disagreement values for the severity assessments is similar to the normal distribution. Removing a few incidents will have no drastic effect on the KA, since most of the values are close to the average and the tails are rather short (the largest value is lower than three times the average). However, if large enough fractions of the raters and incidents are removed, the consensus values do increase to acceptable levels. For example, if KA values above 0.66 are considered acceptable, this can be achieved for the probability assessments by removing 75% of the incidents or 50% of the incidents and 50% of the raters. Considering the severity assessments, acceptable KA values are reached by removing 75% of the incidents, or 50% of the incidents and 25% of the raters, or 25% of the incidents and 50% of the raters, or 75% of the raters. Consequently, it is certainly possible to find subsets of incidents and raters resulting in acceptable KA values. However, none of the combinations identified contains more than 37.5% of the assessed values.

Considering the PCs, removing the same fractions of incidents and raters as for the KA values yields results above the level reached by accounting auditors (0.76) [4]. Although, the consensus of the remaining raters considering the remaining incidents clearly increases when incidents and raters are systematically removed, the fraction of assessments that needs to be removed is quite large (for KA it is over 60%). Thus, the hypotheses H5 and H6 are not supported. The implication is that the low consensus values are not caused by a few exceptionally hard incidents or a few poor raters but rather the rating being difficult in general.

4.4 Limitations

There are several limitations of the study that may have affected the results. The length of the incident descriptions was strictly limited. This was necessary to allow questionnaires with 105 incidents to be constructed and may have led to incidents that were difficult to interpret. However, previously performed test-retests suggest that most respondents understood the questions well enough to provide similar answers when retested [10]. This indicates that the scenarios were comprehendible. Moreover, as quantitative data supporting the assessment of probabilities and severities are rarely available and many security assessments are made in day-to-day work by information security risk assessment experts as well as non-experts, the limited amount of information supporting the assessments may correspond to realistic scenarios.

There are numerous ways that the present study could be extended to incorporate additional aspects. Currently, the level of expertise is based on self-assessments, in future studies other means to grade the expertise of the respondents could be incorporated. Moreover, more advanced optimization algorithms could be used when selecting incidents and raters to be removed, in order to improve the consensus. Further, the study is based on ratings performed by individuals. Additional studies are needed to analyze the consensus when the ratings are performed in group settings. To enable further analysis of the differences between the computed KA values, bootstrapping can be used to decide confidence intervals for the KA values [17].

5 Conclusions

The conclusion of the study is that the assessments of the probability and the severity of information security incidents have to be more stringent than the process used in this study. There is also a need for data supporting the assessments. The lack of underlying data is also supported by the fact that experts and non-intuitive raters do not perform better.

Despite the rather low consensus values, the results do show that there is consensus (inter-rater reliability) among the raters, although not at the desired level. With the support of more underlying data and software tools, the levels of consensus may be increased and turn information security risk assessments into a viable basis for information security management processes.

6 Acknowledgment

This research was funded by the Swedish Civil Contingencies Agency as part of the research program SECURIT (www.foi.se/securit) and the R&D program of the Swedish Armed Forces.

7 References

[1] ISO/IEC, “ISO/IEC 27005:2011 — Information technology — Security techniques — Information security risk management,” 2011.

[2] H. J. Einhorn, “Expert judgment: Some necessary conditions and an example.,” *J. Appl. Psychol.*, vol. 59, no. 5, pp. 562–571, 1974.

[3] H. Holm, T. Sommestad, M. Ekstedt, and N. Honeth, “Indicators of expert judgement and their significance: An empirical investigation in the area of cyber security,” *Expert Syst.*, vol. 31, no. 4, pp. 299–318, 2013.

[4] J. Shanteau, “Why task domains (still) matter for understanding expertise,” *J. Appl. Res. Mem. Cogn.*, vol. 4, no. 3, pp. 169–175, 2015.

[5] A. H. Ashton, “Does Consensus Imply Accuracy in Accounting Studies of Decision Making?,” *Account. Rev.*, vol. 60, no. 2, pp. 173–185, 1985.

[6] J. F. Edens, B. N. Penson, J. R. Ruchensky, J. Cox, and S. T. Smith, “Interrater reliability of Violence Risk Appraisal Guide scores provided in Canadian criminal proceedings,” *Psychol. Assess.*, vol. 28, no. 12, pp. 1543–1549, 2016.

[7] M. H. Epstein, M. K. Harniss, N. Pearson, and G. Ryser, “The Behavioral and Emotional Rating Scale: Test-Retest and Iter-Rater Reliability,” *J. Child Fam. Stud.*, vol. 8, no. 3, pp. 319–27, 1999.

[8] S. Fenz, J. Heurix, T. Neubauer, and F. Pechstein, “Current challenges in information security risk management,” *Inf. Manag. Comput. Secur.*, vol. 22, no. 5, pp. 410–430, 2014.

[9] M. Korman, T. Sommestad, J. Hallberg, J. Bengtsson, and M. Ekstedt, “Overview of Enterprise Information Needs in Information Security Risk Assessment,” *18th IEEE Int. Enterp. Distrib. Object Comput. Conf.*, pp. 42–51, 2014.

[10] T. Sommestad, H. Karlzén, P. Nilsson, and J. Hallberg, “An empirical test of the perceived relationship between risk and the constituents severity and probability,” *Inf. Comput. Secur.*, vol. 24, no. 2, pp. 194–204, 2016.

[11] H. Karlzén, J. Bengtsson, and J. Hallberg, “Assessing Information Security Risks Using Pairwise Weighting,” in *3rd International Conference on Information Systems Security and Privacy, ICISSP 2017*, 2017.

[12] K. Hallgren, “Computing Inter-Rater Reliability for Observational Data: An Overview and Tutorial,” *Tutor. Quant. Methods Psychol.*, vol. 8, no. 1, pp. 23–34, 2012.

[13] M. J. Allen and W. M. Yen, *Introduction to Measurement Theory*. Waveland Press, 2001.

[14] K. Krippendorff, “Computing Krippendorff’s Alpha-Reliability,” *Dep. Pap.*, p. 12, 2011.

[15] M. Lombard, J. Snyder-Duch, and C. C. Bracken, “Content Analysis in Mass Communication: Assessment and Reporting of Intercoder Reliability,” *Hum. Commun. Res.*, vol. 28, no. 4, pp. 587–604, 2002.

[16] D. Freelon, “ReCal OIR: Ordinal, interval, and ratio intercoder reliability as a web service,” *Int. J. Internet Sci.*, vol. 8, no. 1, pp. 10–16, 2013.

[17] A. F. Hayes and K. Krippendorff, “Answering the Call for a Standard Reliability Measure for Coding Data,” *Commun. Methods Meas.*, vol. 1, no. 1, pp. 77–89, 2007.

[18] K. Krippendorff, “Reliability in content analysis: Some common misconceptions and recommendations,” *Hum. Commun. Res.*, vol. 30, no. 3, pp. 411–433, 2004.