Proceedings of the Eleventh International Symposium on
Human Aspects of Information Security & Assurance (HAISA 2017)

How Reliable are Experts' Assessments? A Case Study on UAV Security

A. Shoufan and E. Damiani

Information Security Center-Khalifa University
e-mail: abdulhadi.shoufan|ernesto.damiani@kustar.ac.ae

Abstract

Experts' opinion is a vital source in the information security process. However, the judgement of information security professionals is not always consistent and different experts may provide clearly different ratings. This paper proposes an experimental design towards a quantitative analysis of inter-rater reliability in the field of information security. Twenty experts were asked to rate the security objectives (confidentiality, integrity, and availability) of civilian drone communication in 45 different use cases. Three rates were available: low, medium, and high. The experts' rating was analyzed using Fleiss' kappa to measure the inter-rater reliability. The results show only a slight agreement among the experts which raises concerns regarding the validity of such assessment. However, the experts show higher agreement on the extremes, i.e., when the use case shows clearly high or clearly low security objectives. Increasing the number of experts causes an initial improvement of Fleiss' kappa. However, the latter seems to reach a saturation point when the number of experts exceeds ten, suggesting that large panels do not guarantee increased agreement. Most polled experts seem to have bias towards giving a specific rate. Interestingly, unbiased experts show higher agreement among themselves compared to biased ones. Our findings suggest that the experts' rating should be followed by a verification procedure towards determining the reliability level of the provided data. Also, a purposeful identification of panel subsets with higher inter-rater agreement should be considered.

Keywords

Experts' qualitative assessment, inter-rater reliability, Fleiss' kappa,

1. Introduction

Soliciting expert opinion is a key step in the information security assessment process. Experts are frequently invited to rank or rate vulnerabilities, threats, attacks, assets, risks, and security objectives according to different criteria. For example, Whitman asked information technology executives to rank information security threats, to estimate attack frequencies, and to prioritize expenditures (Whitman, 2003). Loch et al. prepared a list of threats to information systems and resident data from the literature and asked a panel of security consultants and executives to rank the top three (Loch *et al.*, 1992). Experts are also surveyed about security aspects of innovative technologies. For instance, Weiss asked experts to assess privacy concerns in social networks (Weiss, 2009). Expert opinion is also used in the initial stages of event or incidents of potential concern to facilitate rapid risk assessment,

i.e. to give an estimate of risk posed by a threat. Rapid risk assessment is a core part of incident response and thus widely undertaken by security professionals.

Despite its wide usage, polling security experts is not perfect and the reliability of experts' opinion was questioned by several authors. For instance, Halsum *et al.* addressed the uncertainty of risk assessment processes and associated it with the inconsistency in expert judgment (Halsum *et al.*, 2007). Some researchers investigated ways to improve the reliability of experts' ratings as will be discussed in Section 2. However, there has been no quantitative assessment of inter-rater agreement in the field of information security. This paper has two contributions: (i) We propose an experimental design to measure inter-rater agreement among security assessors using Fleiss' kappa statistics. (ii) We validate our design using an extensive case study on a real security problem and analyze its results.

We use the classic CIA triad (Confidentiality-Availability-Integrity) to describe security objectives. Relying on the CIA triad for standardizing security experts' responses is well-documented in the literature and customary practice in the field. For example, CIA triad rating is a standard practice by the National Institute of Standards and Technology (NIST) for security assessment of federal administration's ICT systems (PUB, 2004). As a case study, we investigate the communication security between Unmanned Aerial Vehicles (UAVs or drones) and ground stations. We argue on the one hand that this domain has a strong link to the classic domain of communication security to assure some confidence on the part of polled security professionals. On the other, UAV security is a relatively novel field which should prevent experts from given rates based on "conventional wisdom".

We selected 45 use cases of civil drones and asked a panel of twenty experts to rate the security objectives for each case using the CIA triad. The experts' responses were analyzed using Fleiss' kappa, a specific statistics test for inter-rater reliability. We found out that the experts' overall agreement is low enough to raise serious concerns (below 20 % for all security objectives). However, closer analysis shows a trend to "agree on the extremes": expert agreement is clearly higher when the general perception of the security level of the use case is especially low or especially high.

A further analysis was performed to check experts' bias and how such bias affects the inter-rater reliability. We found out that 14 out of the 20 experts show a permanent tendency to give a low, medium, or high rate, regardless of the case they are assessing. The 20 experts were then divided according to their rating bias and the inter-rater reliability in each subset was analyzed separately. Interestingly, we found out that unbiased raters show better inter-rater reliability. This result hints at using bias control on toy problems as a technique for expert selection. Also, the impact of the number of the experts was investigated. We found out that Fleiss' kappa increases with the number of experts as long as the latter is below 10. Increasing the number of raters beyond 10, however, does not affect the agreement level.

We claim that analyzing the influence of inter-rater agreement on panel-based security ratings can provide some operational suggestions to ensure that these ratings

help rather than harm businesses' security decision-making. According to the principles endorsed by the US Chamber of Commerce for security ratings, reporting expert opinions should "include a coordinated process for adjudicating errors or inaccuracies". Our results suggest that a posteriori analysis of inter-rater agreement should become a key part of such a coordinated process, as well as of other risk assessment procedures that make use of expert ratings.

The remainder of the paper is structured as follows. Section 2 reviews the related work. The methodology is presented in Section 3. Our experiment and its results are described in Section 4 and discussed in Section 5. Section 6 concludes the paper.

2. Related Work

Uncertainty associated with qualitative methods in information security is a well-known issue. Some authors relate it to using imprecise natural language for communication (Smith *et al.,* 2007). While rating risks using ordered categorical labels such as "low", "medium", and "high" can simplify risk assessment, some researchers believe that this approach does not necessarily improve decisions (Cox *et al.,* 2005). To mitigate the impact of uncertainty in qualitative methods, some researchers proposed applying the classic Delphi method to security analysis (Van Deursen *et al.,* 2013). Miller *et al.* attribute the uncertainties of designing secure software systems to missing data on uncommon attacks, difficulty of security cost estimation, and continuous change in technology and tools (Miller et al., 2016). The authors claim that uncertainty has an impact on the experts' perceptions of security risks, which in turn leads to wide variations in their assessments of potential attacks' probability and severity. Their approach is based on Spearman's Rho statistics, which measures the statistical dependence of two sets of rankings. It takes values in the range between *-1* and *+1*, whereas *-1* and *+1* indicate perfect negative or positive correlation, respectively, and *0* indicates no correlation. The authors found rankings of the same attacks across multiple scenarios to be weakly or un-correlated.

In many cases, however, there is no specification of attack scenarios, for example, when rating emerging technologies whose attack surface is still unclear. In such a case, risk assessors only have access to information about potential or actual use cases, with limited insight into underlying technology and real infrastructure, and even less into security controls already in place. Under these conditions, use-case based specification of security objectives can be used as a first step towards categorizing the system and selecting appropriate security controls, as suggested in the Risk Management Framework by the national institute of standards and technology (Stine *et al.,* 2008). In the development of the Guidelines for Smart Grid Cybersecurity, NIST relied on a use-case based approach where the CIA triad security objectives of each use case was rated as low, moderate, or high. However, it is not clear whether this assignment was performed by one or multiple experts (Pillitteri and Brewer, 2014).

For the collection of experts rating data, our work relies on a use case-based approach for drone security where the experts assigned a level to each CIA triad

security objective. This scenario allowed us to study the inter-rater reliability for three security objectives separately: confidentiality, integrity, and availability of information data. We are not aware of any related work that applied inter-rater reliability to experts' assessment data in the context of information security. The choice of drone security is mostly due to the fact that, although highly critical, security of unmanned aerial vehicles has not yet obtained sufficient attention in the research community. We argue that combination of serious security issues and lack of conventional wisdom over them makes the emerging field of drone security the ideal ground where to study security rating.

3. Methodology

Survey development: First, we created a comprehensive list of drone use cases based on media and literature reports. Examples of the listed 45 use cases include climate monitoring, remote sensing, film industry, mineral exploration, volcano monitoring, package delivery, gas pipeline inspection, search and rescue, emergency response, and borderline monitoring. Then, a rating scheme for drone security objectives was developed as given in Table 1. The levels of confidentiality, integrity, and availability for each use case can be rated on an ordinal scale of low, medium, and high for the information data sent from the drone to the ground station.

	High	**Medium**	**Low**
Confidentiality	Data is highly sensitive or has high commercial value	Data is not sensitive and has medium commercial value	Data is not sensitive and has no commercial value
Integrity	Highly critical data with real-time requirements. Data manipulation causes high severity levels	Critical data, however, without real-time demand	Less critical data
Availability	High data rates and real-time response to data content is required	Either high data rates or real-time response to data content are required	Low data rates and time-tolerant response (or no response) to data content is required

Table 1: Rating scheme for CIA security objectives

The rating scheme is explained using some examples. Most commercial drones are deployed for sensor-based applications where different data are collected on the fly and submitted to ground or stored on board for later processing and analysis. The *confidentiality* level of these data tightly relates to the use case. Images submitted by drones on missions for critical infrastructure inspection such as oil and gas pipelines are classified as highly confidential, in general. In some applications, such as library bookshelf monitoring, the data provided by the drone are of less commercial or private value so that its confidentiality can be classified as low. In other use cases such as in the film industry, the producer may prefer to keep the data secret within

limited time frames and the confidentiality level can be classified as medium. While *integrity* is a fundamental requirement for any information, specifying the level of integrity is important to describe the impact of not achieving the corresponding security objective. In many applications, data collected by drones are evaluated or processed on the ground. For example, drones used in rescue operations are sometimes supplied with thermal imaging sensors to allow detection in the night or in invisible areas. Manipulating such sensor data may cause the rescue team to lose sight of injured or trapped people. So, the desired integrity level for such data should be considered high. In some use cases such as climate monitoring, the data integrity is important for accurate simulation or future predictions but there is no real-time requirements for an urgent response. In such cases the integrity level can be classified as medium. In some applications such as road inspection, the drone sends visual image data that are only inspected by the pilot or other operators for uncritical surveillance or inspection purposes. In such cases, the required level of integrity can be described as low. The *availability* level of information data flows essentially depends on the timing requirements for these data. A drone which streams high-rate data to enable a real-time response such as in the case of Emergency Response should show high of availability. When the data rate is low or the response time is not especially critical we classify the availability level as medium such as in the case of Ship Inspection. In use cases, where neither high data rate are required nor a real-time response to the data is expected, e.g. in remote sensing, the data availability level can be classified as low.

Data collection: Twenty experts from five different countries in Asia, Europe, and the USA were addressed individually and briefed about the purpose of the study and its methodology. Then a document was sent to each expert, including the rating scheme given in Table 1, its explanation as given above, and a table with all use cases. The experts were asked complete the table by rating each CIA component for each use case as low, medium, or high.

Determination of inter-rater agreement: To evaluate the reliability of experts' assessment, Fleiss' kappa statistics was used. This statistic is a chance-corrected measure of agreement among multiple raters. Higher values of Fleiss' kappa are assumed to indicate higher agreement. For space reasons, the reader is referred to (Fleiss *et al.,* 1969) for more details on calculation of this statistics. A widely used, although not generally accepted, interpretation of Fleiss' kappa was provided in (Landis and Koch, 1977) as summarized in Table 2.

Fleiss' kappa	**Interpretation**
<0	Poor agreement
0.01-0.20	Slight agreement
0.21-0.40	Fair agreement
0.41-0.60	Moderate agreement
0.61-0.80	Substantial agreement
0.81-1.00	Almost perfect agreement

Table 2: Fleiss' kappa interpretation according to (Landis and Koch, 1977)

4. Results

Fleiss' kappa was computed for each CIA security objective. Table 3 summarizes the results which indicate a slight agreement according to Landis and Koch's scheme given in Table 2. While there is no agreement on acceptable values of Fleiss' kappa in security ratings, we remark that in some domains, agreement levels below "fair", i.e. "poor" and "slight", cause immediate rejection of the rating (Everitt, 1992).

Security Objective	Fleiss' kappa
Confidentiality	13.8%
Integrity	15.5%
Availability	19.1%

Table 3: Fleiss' kappa for the assessed security objectives

To understand in which use cases assessors showed lower or higher agreement, we studied the observed agreement for each use case as a function of a new metric we called Security Objective Index (SOI). The observed agreement is the pure agreement found in the rating data without a "by-chance component". We refer the reader to (Fleiss *et al.,* 1969) for more details on the calculation of this statistic. SOI reflects the experts' overall perception of the level of some security objective for some use case. It is defined as $SOI= (n_L+2n_M+3n_H)/(n_L+n_M+n_H)$, where n_L, n_M, and n_H refer to the numbers of low, medium, and high rates given to the security objective of some use case, respectively. The value of SOI varies between 1 (when all raters assigned low) and 3 (when all raters assigned high). Then, for each of the three security objectives, the use case observed agreement was plotted against the respective security objective index as shown in Fig. 1. Interestingly, in all the three plots a parabolic trend can be observed. This indicates that the observed agreement tends to increase when the security objective index of a use case is close to its extreme values. In other words, when a use case has a high or a low SOI, the assessors tend to show more agreement in their ratings.

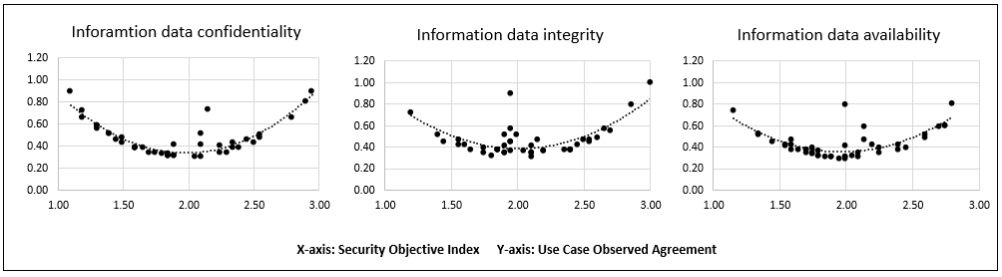

Figure 1: Observed agreement vs. security objective index

To understand whether some experts tend to give a rate more frequently than the other two rates, we determined the total numbers of low, medium, and high rates given by each expert for all use cases and all security objectives. We found out that 14 experts show some bias in their ratings. In particular, three, seven, and four experts tended to give low, medium, or high rates, respectively. The remaining six

raters did not show a "favourite" rate. Thus, a total of four subsets could be identified. Additional tests were performed to determine the inter-rater agreement (Fleiss' kappa) within each subset and for the three security objectives. Then, the three Fleiss' kappa values within each subset were averaged as summarized in Table 4. Obviously, the "unbiased" experts showed the best average inter-rater reliability in contrast to those who tended to give high rates.

Subset	**Mean Fleiss' kappa**
Experts without clear bias	19.6%
Experts biased to give low rates	18.0%
Experts biased to give medium rates	16.2%
Experts biased to give high rates	15.3%
All experts	16.1%

Table 4: Average Fleiss' kappa for different bias subsets

Finally, to investigate the impact of the number of raters on the inter-rater agreement, 10,000 tests were performed for each security objective. In each test a random combination of raters was selected, whereas the combination size is chosen randomly between two and twenty (Note that there is only one combination of size 20). For each combination, Fleiss' kappa was determined. At the end, the obtained kappa values for each combination size, i.e., for each number of raters were averaged. Figure 3 shows the results of these tests. Accordingly, the inter-rater agreement increases with the number of raters until the latter reaches approximately ten. Afterwards, the inter-rater agreement remains almost constant.

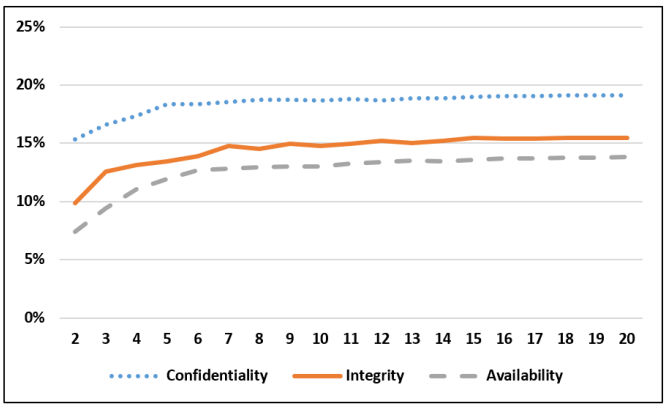

Figure 2: Fleiss's kappa as a function of the number of raters

5. Discussion

The first finding of this study is that the inter-rater reliability among security experts is at best slight when measured by Fleiss' kappa and interpreted according to (Landis and Koch, 1977). This result quantitatively confirms the concerns about the judgment reliability of information security experts raised in the literature, e.g., (Halsum *et al.*, 2007), (Van Deursen *et al.,* 2013), and (Miller et al., 2016). This

suggests that standard security processes involving expert polls should pay more attention to inter-rater agreement. Fleiss *et al.* commented that "if agreement among the raters is good, then there is a possibility, but by no means a guarantee, that the ratings do in fact reflect the dimension they are purported to reflect. If their agreement is poor, on the other hand, then the usefulness of the ratings is severely limited, for it is meaningless to ask what is associated with the variable being rated when one cannot even trust those ratings to begin with" (Fleiss *et al.*, 2013).

While it is important to understand the reasons for the slight agreement shown in this study, it is very difficult to provide a general explanation. However, some remarks can be offered. For example, the level of expertise in drone security seems not to affect the level of agreement: Two of the twenty raters involved in this study are renowned experts in drone security. We calculated the agreement between these two experts and compared it with the average agreement between any other two experts. Interestingly, the agreement between the two drone experts was higher only when they rated the integrity level of information data. For the other two security objectives, the drone experts showed a lower level of agreement. Another factor relates to this type of study in general. Specifically, obtaining experts' assessment is not an easy task especially when it is done on a voluntary basis. The lack of agreement could be related to an imbalance between the interest in the study and the overhead of completing the survey. Remember that the experts first need to study the rating scheme given in Table 1 and read the provided explanation. Then they need to go use case by use case and create an idea about how each use case may look like in reality before they estimate its security objectives. Collecting experts' judgment in terms of rates (here, low, medium, and high) is less informative than a narrative assessment where experts justify their judgment and explain how they pictured a use case in their mind. However, narrative assessment would demand higher time investment from the experts, which cannot be expected when many scenarios or use cases are to be rated like in the present study.

Fleiss' comment cited above points to the difference between agreement and validity of experts' judgments and indicates that a good agreement level is a necessary but not sufficient requirement for the validity of the rating. However, what does validity mean in our study? Assessing validity requires a golden standard to compare assessments with, which is obviously not available in our case. Usually, a vendor, a user, or a third party is interested in using the CIA triad to assess drone security objectives in one or a limited number of use cases, to select and provide appropriate security controls. Thus, the experts' judgment for that specific use case is the most interesting outcome for the user. When the experts' ratings for a specific use case are concentrated around the high or the low value, the user can be more confident about the rating's validity because security experts seem to show high agreement for use cases where the security objectives are obviously high or obviously low according to Figure 1. In contrast, if the experts' ratings are widely distributed over the three rate values, then the result is less useful. A more in-depth analysis of the use case should be conducted by or on behalf of the user to improve confidence in the assessment.

Clustering raters according to their bias and analyzing the inter-rater reliability of the different clusters allow other interesting remarks, although the cluster sizes are too small for generalization. The first interesting observation is that the six raters who assigned low, medium, and high rates without a clear pattern, showed the best average inter-rater reliability, even when compared to panel-wide data (Table 4). This finding is interesting because our intuition suggested otherwise. This can be understood better when compared with biased raters, e.g., the seven panellists who tended to give "medium" rates. Take an arbitrary use case, for example, and ask the medium-biased experts to rate it. Although the medium-biased experts are more probably to rate a security objective as medium, they would show less agreement than unbiased experts.

An essential question in qualitative security studies is the number of experts that should be involved in rating. The selected experts can be considered as a sample of the experts' population. Following basic statistical assumption, a higher sample size is desired. Our study, however, shows that the average inter-rater reliability does not improve, but also does not worsen, when the number of raters exceeds 10 according to Figure 2. We believe that this finding is of general value and we are not aware of any related work that studies the impact of number of experts on the inter-rater reliability. The initial increase in Fleiss' kappa with the number of raters may appear to be counter-intuitive, because we tend to expect less agreement when more raters are involved. This intuition, however, seems to be incorrect when we give the experts a set of rates to choose from. To explain this assume that two raters are asked to rate some item as good or bad. The chance that they agree is 50%. If we add a third rater the chance that the three agree is just 25%. However, the chance that two of three agree is now 75%. Fleiss' kappa considers both complete and partial agreement.

6. Conclusion

Collecting and interpreting experts' assessment is an essential step of many information security processes and practices, including standard procedures for security ratings. This paper described an experiment showing the inter-rater agreement of security experts using the CIA triad to assess a set of operational components of drones. The observed inter-rater agreement on each specific use case, however, increases when the security index is especially low or high. Increasing the number of raters beyond ten does not affect the inter-rater reliability. Also, 70% of experts showed a clear bias toward giving a low, medium, or high rate. Unbiased raters showed better inter-rater agreement in general. Besides the results and the new perspectives discussed above, we believe that this work will pave the way to extensive experimental analysis of the security rating process.

Acknowledgment

This research was funded by the UAE Telecommunications Regulatory Authority. Project: Grid-Enabled Business Process Management and e-Government (ICT fund).

7. References

Bornstein, B.H. and Greene, E., 2017. *The jury under fire: Myth, controversy, and reform*. Oxford University Press.

Cai, G., Dias, J. and Seneviratne, L., 2014. A survey of small-scale unmanned aerial vehicles: Recent advances and future development trends. *Unmanned Systems*, *2*(02), pp.175-199.

Cox, L.A.T., Babayev, D. and Huber, W., 2005. Some limitations of qualitative risk rating systems. *Risk Analysis*, *25*(3), pp.651-662.

Everitt, B.S., 1992. The analysis of contingency tables (monographs on statistics and applied probability 45).

Fleiss, J.L., Cohen, J. and Everitt, B.S., 1969. Large sample standard errors of kappa and weighted kappa. *Psychological Bulletin*, *72*(5), p.323.

Fleiss, J.L., Levin, B. and Paik, M.C., 2013. *Statistical methods for rates and proportions*. John Wiley & Sons.

Landis, J.R. and Koch, G.G., 1977. An application of hierarchical kappa-type statistics in the assessment of majority agreement among multiple observers. *Biometrics*, pp.363-374.

Loch, K.D., Carr, H.H. and Warkentin, M.E., 1992. Threats to information systems: today's reality, yesterday's understanding. *Mis Quarterly*, pp.173-186.

Miller, S., Wagner, C., Aickelin, U. and Garibaldi, J.M., 2016. Modelling cyber-security experts' decision making processes using aggregation operators. *Computers & Security*, *62*, pp.229-245.

Pillitteri, V.Y. and Brewer, T.L., 2014. Guidelines for smart grid cybersecurity. *NIST Interagency/Internal Report (NISTIR)-7628 Rev 1*.

PUB, F., 2004. Standards for Security Categorization of Federal Information and Information Systems.

Smith, G.R., Scouras, J. and DeBell, R.M., 2007. Qualitative representation of risk. *Wiley Handbook of Science and Technology for Homeland Security*.

Stine, K.M., Kissel, R., Barker, W.C., Lee, A., Fahlsing, J. and Gulick, J., 2008. SP 800-60 Rev. 1. Volume I: Guide for Mapping Types of Information and Information Systems to Security Categories; Volume II: Appendices to Guide for Mapping Types of Information and Information Systems to Security Categories.

Van Deursen, N., Buchanan, W.J. and Duff, A., 2013. Monitoring information security risks within health care. *computers & security*, *37*, pp.31-45.

Weiss, S., 2009. Privacy threat model for data portability in social network applications. *International journal of information management*, *29*(4), pp.249-254.

Whitman, M.E., 2003. Enemy at the gate: threats to information security. *Communications of the ACM*, *46*(8), pp.91-95.