Assessing Information Security Risks using Pairwise Weighting

Henrik Karlzén, Johan Bengtsson and Jonas Hallberg
Department for Information Security and IT Architecture, Swedish Defence Research Agency,
Olaus Magnus väg 42, Linköping, Sweden
{henrik.karlzen, johan.bengtsson, jonas.hallberg}@foi.se

Keywords: Risk Assessments, Pairwise Weighting, Information Security Risk, Cognitive Style, Cognitive Load.

Abstract: In practice, assessing information security risks is difficult since available methods lack specificity on how to perform the assessments as well as what input should be used. Thus, the process becomes resource demanding with fairly large rater-dependency. An established way of facilitating rating processes is to weight objects against each other, rather than rating each object independently on an absolute scale. In this paper, we investigate whether such a method, inspired by the Analytic Hierarchy Process, can be useful for assessing information security risks. However, the new method did not result in higher inter-rater reliability or lower cognitive load. This result was true both for experts and non-experts, as well as among raters with different cognitive styles.

1 INTRODUCTION

A large number of methods have been proposed for the assessment of information security risks associated with threats. However, these methods do not provide any substantial guidance on how to perform the underlying assessments of *probability* (likelihood) and *severity* (impact or consequence) or a common description of what input should be used during such assessments (Korman et al., 2014). Unfortunately, this results in assessing probability and severity being difficult in practice (Fenz et al., 2014), with high resource demands and often insufficient reliability among different raters, leading to rater-dependent assessments.

Although rater-independence does not indicate assessments closer to the truth per se, the objective truth typically remains elusive and inter-rater reliability is a suitable surrogate indicator, since it is necessary for validity (Gwet, 2014). For these reasons, it would be useful to find a new way of assessing risks that shows higher inter-rater reliability and results in lower cognitive load, without increasing the number of raters. A potential method would be to compare the threats rather than make an absolute assessment about each one, e.g. similar to the weighting used in the Analytic Hierarchy Process (AHP) (Saaty, 1990). This paper investigates the possible advantages of this approach over the more traditional ones where each threat is assessed independently.

While improved inter-rater reliability is the main goal of applying a new method to risk assessments, there must also be a balance with the required resources to use the methods. In the field of cognitive load theory relating to learning, three types of cognitive processing are included (Deleeuw and Mayer, 2008): the intrinsic processing which relates to the inherent complexity of the task; the extraneous processing which concerns the redundant information included and therefore the presentation; and the germane processing which relates to knowledge and learning. Hence, extraneous load should be kept at a minimum whereas germane load is encouraged in learning situations. However, separating the two in measurements has proven difficult. Measuring cognitive load is typically done with self-ratings (Paas et al., 2003) where participants put numerical values on their own perceived mental burden.

The following hypotheses were tested:

H1. Inter-rater reliability is higher when rating probability using pairwise weighting rather than the traditional method.

H2. Inter-rater reliability is higher when rating severity using pairwise weighting rather than the traditional method.

H3. Cognitive load is lower when rating probability and severity using pairwise weighting rather than the traditional

Karlzén, H., Bengtsson, J. and Hallberg, J.
Assessing Information Security Risks using Pairwise Weighting.
DOI: 10.5220/0006138203180324
In *Proceedings of the 3rd International Conference on Information Systems Security and Privacy (ICISSP 2017)*, pages 318-324
ISBN: 978-989-758-209-7

method. Specifically:

i. Mental effort is lower when rating probability and severity using pairwise weighting rather than the traditional method.

ii. Difficulty is lower when rating probability and severity using pairwise weighting rather than the traditional method.

iii. Time consumption is lower when rating probability and severity using pairwise weighting rather than the traditional method.

Section 2 of the paper describes the method. Section 3 gives the results and section 4 discusses these results.

2 METHOD

The sections below describe the participants, the survey instrument, and the data collection procedure.

2.1 Participants

The survey was distributed to a strategic sample of 10 researchers active in the areas of information security, IT security, IT management or human factors and so were not all experts on risk assessments or information security. All respondents were from the Swedish Defence Research Agency, possess university degrees, and work as researchers, consultants or in management.

2.2 Material and Scales

The study was conducted using two paper-based questionnaires, each questionnaire comprising of three parts which were filled out by each participant. The first part consisted of eight questions about the respondent, and were identical between the two questionnaires, although the respondents only had to answer them on whichever questionnaire they filled out first. The second part consisted of 105 potential incidents that the respondent was asked to assess regarding both probability and severity using visual analogue scales. One questionnaire applied the traditional method, while the other questionnaire instead used weighting. A third part, identical to both questionnaires, concerned cognitive load of filling out the questionnaires.

2.2.1 Incidents

The 105 potential incidents were reused from a previous study which investigated whether people truly tend to perceive risk as a multiplicative function of probability and severity (Sommestad et al., 2016). The incidents were intended to be relatable for the participants and to cover the entire risk matrix, although naturally with fewer incidents with both high probability and severity. Some examples include:

- “A security flaw in the authentication tokens allows a malicious outsider access to the local network.”
- ”An employee installs freeware that covertly copies local and networked folders to a server controlled by a large defence corporation”.
- “An employee gathers large amounts of secret documents concerning IT-security and hands them over to a foreign nation.”

2.2.2 Probability and Severity

The questionnaire for the traditional method asked respondents to rate severity and probability of each incident on each of two lines. Ten helping markers per line were present but exact indicated values were measured using a ruler. The S*everity* scale stretched from 0 (Minimal, no harm at all) to 10 (Greatest harm among all listed incidents). The *probability* scale stretched from 0% (Minimal, completely unlikely for the next ten years) to 100% (Maximal, guaranteed to happen).

Conversely, the second questionnaire compared incidents with each other. The first incident was pitted against the second, the second against the third, the third against the fourth, and so on. For each comparison, the respondents were asked how the incidents compared in both probability and severity separately. A scale was provided where circling the suitable number on the left part indicated that the first mentioned incident had greater probability (or severity if that was measured), while circling the middle 1 meant that the incidents were equal in this regard, and finally circling the suitable number on the right part of the scale indicated that the second mentioned incident had greater probability or severity. The scale ran from a factor of 9 and every odd number: (9, 7, 5, 3, 1, 3, 5, 7, 9), which corresponds to the commonly used original scale in AHP (Ishizaka and Labib, 2011). The numbers on the scale were also explained as: 1 – equal, 3 – moderately more important, 5 – strongly more important, 7 – very strongly more important, and 9 – extremely more important. Since each comparison introduced one new incident (except the first one which introduced two), a total of 104 comparisons were needed for probability and severity respectively in order to enable the computation of weights for each of the 105 incidents (Ishizaka and Lusti, 2004).

2.2.3 Cognitive Load

To measure the impact on the participants when assessing, three aspects of cognitive load were gauged on each of the two questionnaires.

One question asked about the mental effort to fill out the questionnaire (Likert scale 1–7 from extremely low to extremely high effort). This is similar to the effort scale in e.g. (Paas, 1992) and is related to intrinsic cognitive load (Deleeuw and Mayer, 2008). Another question concerned how difficult it was to fill out the questionnaire (Likert scale 1–7 from extremely simple to extremely difficult), similar to e.g. (Marcus et al., 1996) and relating to germane cognitive load, according to (Deleeuw and Mayer, 2008).

As suggested in (Marcus et al., 1996), both self-reported rating scales were administered immediately after the main test as to ensure that evaluations are fresh in memory. Both self-ratings have been shown to be reliable, sensitive and do not impact performance (Paas et al., 1994).

Furthermore, the time it took each respondent to fill out the questionnaire was (objectively) measured. This factor is sometimes underestimated (Paas et al., 2003) although it was measured in e.g. (Fink and Neubauer, 2001). In (Deleeuw and Mayer, 2008) response time (to a concurrent secondary task) related to the third cognitive load dimension of extraneous load.

2.2.4 Cognitive Style and Expertise

Eight questionnaire items related to decision making and were taken from (McShane, 2006). Four of these measured rationality tendency with a focus on objective information and logical analysis. The other four instead measured the respondents’ propensity to utilise intuition and instinct rather than rationality.

Three further items related to expertise concerning information security, IT security, and risk assessments.

2.3 Data Collection

A crossover study with counterbalancing was used. To alleviate order effects, the respondents were randomly divided into two equal groups. One group assessed the items using the traditional method for a first session and used the new weighting method for the second session. The reverse order was applied for the other group. General risk analysis learning effects should be fairly equal between the two methods, so an order effect is unlikely and indeed no statistically significant order effect was found.

To make sure the respondents’ assessments on the second questionnaire were not affected by the first, there was a gap of at least one week between questionnaires. Also, the respondents were instructed not to keep notes and were not told the specific aim of the study beyond the investigation of risk perceptions.

2.4 Internal Validity Measurement

2.4.1 Probability and Severity

We constructed our own incidents, which may have led to incidents that were difficult to interpret. However, we are primarily interested in comparing the reliability of two methods, which is fairly robust against ambiguity of incidents, especially in view of the fairly large number of incidents. As will be seen, incidents were clearly well enough understood.

Also, the answers for each respondent showed correlations between probability and severity in (-0.710)–(-0.219) for the traditional method. A negative correlation is natural since most incidents have high values for at most one of probability and severity, and is in line with e.g. -0.56 in (Weinstein, 2000).

2.4.2 Cognitive Load

Standardised Cronbach's Alpha = 0.797 (95 % CI 0.573–0.913) showing acceptable reliability, i.e., they were internally consistent, indicating that the three items measure the same construct of cognitive load, so we do not see different types of cognitive load. The situation in the literature is not clear, some studies gives support for our model, others see distinct types of cognitive load, e.g. (Deleeuw and Mayer, 2008). For instance, they have a statistically significant correlation of only 0.33 between effort and difficulty in one experiment compared to our 0.595 ($p < 0.01$). This is reasonable seeing as higher complexity demands more effort, although effort can be high regardless and increased effort cannot handle all increases of complexity.

2.4.3 Cognitive Style and Expertise

The eight cognitive style items had a Cronbach's Alpha = 0,913 (95 % CI 0.797–0.975) showing very high internal consistency, which is consistent with the item basis.

The expertise items had a Cronbach's Alpha = 0,756 (95 % CI 0.286–0.934) where removing question 9 (working with security or risk assessments) would produce considerably higher alpha, expectedly suggesting that this reflects a separate construct from questions 10 and 11 (which are about working with security more generally). It should be noted that self-ratings of expertise can be ambiguous, since more knowledgeable people can tend to be more humble about their abilities, e.g. in (Holm et al., 2014).

3 RESULTS

3.1 Inter-rater Reliability

3.1.1 Probability

Inter-respondent reliability for probability using the traditional method had Cronbach’s alpha of 0.861 (95 % CI 0.817–0.897) with corrected item-total correlations 0.358–0.714.

For the new method, each item asked for a rating pitting two incidents against each other and, to homogenise each rater’s scale, ratings were transposed to express each incident in terms of the first incident, I12. Since the value of each rater’s I12 was not gaged, each of the rater’s ratings may depend on different I12s, resulting in a multiplicative effect. To eliminate such a possible effect, each rater’s ratings were independently standardised. Inter-rater reliability for probability using the new method had standardised Cronbach’s alpha of 0.805 (95 % CI 0.744–0.857) with corrected item-total correlations -0.111–0.804 where three of the raters were below 0.3. So, the traditional method performed better in terms of error for probability, although for the raters as a whole it was only a slight difference with the confidence intervals overlapping in part.

3.1.2 Severity

Inter-respondent reliability for severity using the traditional method had Cronbach’s alpha of 0.908 (95 % CI 0.880–0.932) with corrected item correlations 0.491–0.818. Inter-rater reliability for severity using the new method had standardised Cronbach’s alpha of 0.415 (95 % CI 0.232–0.569) with corrected item correlations 0.080–0.260. So, the traditional method performed much better in terms of error for severity, with the new method having very low reliability.

3.2 Cognitive Load

3.2.1 Mental Effort

Self-reported mental effort was on average 1.3 points higher for the new method, which is a large effect size (absolute value of Cohen’s d = 0.80, p = 0.022 < 0.05). So hypothesis H3.i was not supported.

3.2.2 Difficulty

Difficulty ratings were on average 0.6 higher for the new method, equivalent to a small effect size (absolute value of Cohen’s d = 0.32), so no support for hypothesis H3.ii, although this is not statistically significant (p = 0.382).

3.2.3 Time

The questionnaire for the new method took on average 15.8 minutes longer than the old method, equivalent to a large effect size (absolute value of Cohen’s d = 0,94, p = 0.004 < 0,05). Hence, hypothesis H3.iii was not supported.

4 DISCUSSION

4.1 Hypotheses

While it is not possible to know whether the raters’ ratings reflect true probability and severity of the incidents, the new method performed worse in regard to all measured factors: the inter-rater reliability for probability (overlapping CIs) and severity, time, mental effort, and difficulty (although not statistically significantly for the last factor). Hence, none of the hypotheses were supported.

Furthermore, it is important to keep in mind that measuring probability and severity is usually a stepping stone to estimating risk. Since risk is the product of probability and severity, overall reliability will be at most as high as the lowest reliability of the constituents, cf. (Krippendorf, 2004).

4.2 Increasing Reliability and the Role of Cognitive Style and Expertise

To improve reliability, any differences between raters, including cognitive style and expertise, must be addressed.

Raters may let personal feelings and attitudes towards the outcomes (severity) of the incidents play a role, e.g. not caring that the organisation loses a document since the rater does not really care about the organisation, or being overly risk-averse and easily scared. This amounts to a systematic difference between raters. Cronbach’s alpha treats systematic inter-rater differences as irrelevant and are equivalent to intra-class coefficients (ICC) for consistency. Not ignoring systematic differences and thus using ICC for absolute agreement, the coefficients decrease by approximately 0.05 for each of probability and severity using the traditional method. This shows that systematic error is not a large part of the reliabilities, but can nevertheless be meaningful to target for an improving organisation. It should be noted that standardising scores removes systematic differences so no similar calculation can be performed for the new method. An additive difference between raters would however skew the calculations for the new method, since each transposed score in terms of the first incident would be on the form:

$$(x_i + a) / (x_{i-1} + a) \quad (1)$$

rather than simply:

$$x_i / x_{i-1} \quad (2)$$

With a small overall systematic difference of -0.05 this should however not be a major issue. Furthermore, in practice, starting questionnaires with calibration of each rater’s responses would alleviate this.

Furthermore, raters may have different knowledge about the incidents, e.g. what use an attacker can make of a stolen document, or raters may not be very used to rating information security incidents, at least with the specific method. Raters may also differ in how used they are to risk analysis and logical thinking. Item 7 on cognitive style was correlated (0.733, p = 0.016) with probability for the new method (in terms of corrected item-total correlation), implying that raters with a more logical cognitive style showed more inter-rater reliability with other raters. On the other hand, items 2 and 5 were just about correlated (-0.627, p = 0.053 and -0.629, p = 0.051) with severity of the traditional method, implying that raters with a more logical cognitive style showed less inter-rater reliability with other raters. There were no other statistically significant correlations concerning probability or severity and cognitive style or expertise. All in all, these results do not show any clear relationship between cognitive style or expertise and inter-rater reliability, for either method. This is not entirely surprising since experts do not typically perform better in areas where systems are dynamic and behavioural, with limited outcome feedback, as is the case in information security (Shanteau, 2015). It is also feasible that the task is more related to other fields than information security, such as business sense or systems engineering.

4.3 Possible Limitations

In contrast with e.g. the NASA Task Load Index (TLX) (Luca, 2014) we do not measure cognitive load in terms of physical effort, but this should anyway be very low for our study. Likewise, we do not explicitly measure TLX’s frustration or the Subjective Workload Assessment Technique’s stress factor (Luca, 2014). However, the exact scale – or even whether it is unidimensional or multidimensional – and the use of verbal labels, is not critical in measurements of cognitive load (Sweller et al., 1998) and both frustration and stress would intuitively seem to map to effort ratings with no further fine-grained measures necessary here.

Another possibly limiting factor is the length of the questionnaires. Comparing the inter-rater reliability between the first 52 items and the remaining 52 or 53 items of the questionnaire for each method, showed practically no difference for the traditional method’s probability and a small difference for the traditional method’s severity (alpha 0.929 for the first half compared to 0.881 for the second), displaying very little impact of the length of the questionnaire. This is fortunate, because 105 incidents is not likely an unusually large amount to rate in one sitting. For the new method, the inter-rater reliability for probability was much lower for the first half (0.303 vs. 0.862), while severity showed the reverse with a higher first half (0.581 vs. 0.298). As the response on each item on the new method depends on all previous items, it is unsurprising that the new method produces large differences over time, and the improvement for the second half of probability is likely inflated because of this. In fact, closer examination shows that the split data no longer fits the necessary underlying models for the new method.

5 CONCLUSIONS

All in all, information security risk assessments using the method based on pairwise weighting tested in this paper cannot be recommended. However, before dismissing pairwise weighting altogether, there are a few possible modifications to be evaluated. First, the use of the traditional AHP scale for the comparisons should be compared to the merits of using other scales, such as scales based on fewer steps or different sets of values assigned to the steps of the scale.

Secondly, alternative approaches to selecting the pairs of threats to be compared should be tested. Ideally each pair of threats should be compared. However, such an approach would be highly cumbersome to the raters since the number of necessary comparisons grows by roughly the total amount of threats for each additional threat. Conversely, the approach used in this study is based on the lowest possible number of comparisons, which although less unwieldy cannot easily account for inconsistencies in inter-respondent ratings. Redundancy in the comparisons could be used to decrease the problem of inconsistent weightings and provide an overall more consistent results among the respondents. A probable improvement would be to utilise a software tool to give raters a better overview of the threats as a whole, while also facilitating backtracking and further analysis.

Consequently, there is room for more experiments on using pairwise weighting for information security risk assessments.

ACKNOWLEDGEMENTS

This work was conducted in the FOI research project Assessment and Analysis of IT Systems, which is funded by the R&D program of the Swedish Armed Forces.

REFERENCES

Deleeuw, K. & Mayer, R., 2008. A Comparison of Three Measures of Cognitive Load: Evidence for Separable Measures of Intrinsic, Extraneous, and Germane Load. *Journal of Educational Psychology*, Vol. 100, No. 1, 223–234.

Fenz, S., Heurix, J., Neubauer, T., & Pechstein, F., 2014. Current challenges in information security risk management. *Information Management & Computer Security*, 22, 410–430.

Fink, A., & Neubauer, A., 2001. Speed of information processing, psychometric intelligence, and time estimation as an index of cognitive load. *Personality & Individual Differences*, 30, 1009–1021.

Gwet, K. L., 2014. Handbook of Inter-Rater Reliability: The Definitive Guide to Measuring The Extent of Agreement Among Raters (4th ed.). *Advanced Analytics*, LLC.

Holm, H., Sommestad T., Ekstedt M., & Honeth, N., 2014. Indicators of expert judgement and their significance: an empirical investigation in the area of cyber security. *Expert Systems*. Volume 31, Issue 4, pages 299–318.

Ishizaka, A., & Labib, A., 2011. Review of the main developments in the analytic hierarchy process. *Expert Systems with Applications*, 38(11), 14336–14345.

Ishizaka, A., & Lusti, M., 2004. An expert module to improve the consistency of AHP matrices. *International Transactions in Operational Research*, 11(November), 97–105.

Korman, M., Sommestad, T., Hallberg, J., Bengtsson, J., & Ekstedt, M., 2014. Overview of Enterprise Information Needs in Information Security Risk Assessment. Proceedings of the 18th IEEE International Enterprise Distributed Object Computing Conference (EDOC). pp. 42-51.

Krippendorff, K., 2004. Reliability in content analysis: Some common misconceptions and recommendations. *Human Communication Research*. Vol. 30, pp. 411-433.

Luca, L., 2014. Formalising Human Mental Workload as a Defeasible Computational Concept. *A Dissertation submitted to the University of Dublin*, Trinity College.

Marcus, N., Cooper, M., & Sweller, J., 1996. Understanding Instructions. *Journal of Educational Psychology*. Vol. 88, No. 1, 49-63.

McShane, S., 2006. Activity 8.8: Decision Making Style Inventory. In *Canadian Organizational Behaviour*. McGraw-Hill Education.

Paas, F., 1992. Training strategies for attaining transfer of problem-solving skill in statistics: A cognitive-load approach. *Journal of Educational Psychology*, 84, 429–434.

Paas, F., Tuovinen, J., Tabbers, H. & Van Gerven, P., 2003. Cognitive Load Measurement as a Means to Advance Cognitive Load Theory. *Educational Psychologist*, 38(1), 63–71.

Paas, F., van Merriënboer, J., & Adam, J., 1994. Measurement of cognitive load in instructional research. *Perceptual and Motor Skills*, 79, 419–430.

Saaty, T. L., 1990. How to make a decision: The analytic hierarchy process. *European Journal of Operational Research*, 48(1), 9–26.

Shanteau, J., 2015. Why Task Domains (Still) Matter for Understanding Expertise. *Journal of Applied Research in Memory and Cognition*, July 2015.

Sommestad, T., Karlzén, H., Nilsson, P., & Hallberg, J., 2016. An empirical test of the perceived relationship between risk and the constituents severity and probability. Information & Computer Security. Volume 24, Issue 2.

Sweller, J., van Merriënboer, J., & Paas, F., 1998. Cognitive Architecture and Instructional Design. *Educational Psychology Review*, Vol. 10, No. 3.

Weinstein, N., 2000. Perceived probability, perceived severity, and health-protective behavior. *Health psychology : official journal of the Division of Health Psychology*, American Psychological Association, 19(1), pp.65–74.