On the Sensitivity of Cyber Assessment Methodologies

Paul D. Rowe, J. Cory Miniter, Michael D. Norman

The MITRE Corporation

Abstract

A common feature of cyber security and resiliency assessment methodologies is to elicit semi-quantitative information from subject matter experts (SMEs). This information is frequently based on expert knowledge of the capabilities and motivations of hybrid threats. SMEs typically provide ratings of various system aspects on an ordinal (e.g., 1-5) scale which is then aggregated to create a prioritized rank order. Crucial system information may be hidden or lost during such assessments. Here we present an approach which is cognizant of multiple sources of complexity that exist in SME-driven cyber resiliency assessment methodologies.

Introduction

Various applied research institutions, such as MITRE, have developed and implemented multiple methodologies for assessing the security and resiliency of large cyber systems. Attempts to sample and model systems are often themselves influenced by the complexity of human psychology (Liebovitch et al., 2011). The complexity we observe during sampling is the result of human decision-making. We can assume that each individual's interpretation of an ordinal rating scale (e.g., 1-5) at each sampling event has the potential to be nonlinear. Understanding where these nonlinear cases really matter is ultimately the goal of investigating the sensitivity of cyber assessment methodologies.

In order to confront this challenge, one of the key objectives for this work is to put forth a novel method for capturing more detail about the decision-maker's perspective during the

assessment elicitation process. Another key objective of this work is to suggest a repeatable method for using such details to model the system under scrutiny a bit more accurately.

Methodologies used to assess cyber resiliency include Adversary-Driven Cyber Resiliency (ACR) (McQuaid et al., 2017), Crown Jewels Analysis (CJA) (Watters and Morrissey, 2018), and Threat Assessment and Remediation (TARA) (Wynn et al., 2011), among numerous others. While these methodologies have slightly different approaches and goals, they share some features. For one, they were all developed with advanced hybrid threats in mind. Their purpose is to aid decision makers in planning for mitigations against the most harmful effects posed by adversaries capable of working across the full spectrum of domains, from supply chain attacks to influence campaigns. Another technical commonality among these methodologies is that, after identifying a set of system features, they typically include a process of elicitation of subject matter expert (SME) input to rank these features along various dimensions, with the end goal of creating a single rank-ordered list to guide decision makers in how to spend valuable and limited resources to improve system resiliency.

A straightforward way of creating such a rank-ordered list of features is to present the SMEs with the set of features to be ranked, and have them directly create the rank ordered list. Although it is simple to describe, this method is fraught with problems. For one, it is dominated by subjectivity. Two SMEs are likely to give different responses. Indeed, the same SME may be prone to giving different responses on different days. This makes it diffcult for a decision maker to act on these results. The risk posed by an advanced, hybrid threat requires the ability of the decision maker to probe the complex assumptions hidden behind a single number encoding expert judgment.

In attempting to deal with this human-induced complexity, a common approach employed by various assessment methodologies (see Figure 1) is to elicit from the SMEs various Likert-scale (e.g. 1-5) ratings of the system features along various dimensions. For example, ACR asks SMEs to rate various potential attacks against a system according to dimensions of diffculty (e.g. diffculty of access or of command and control) and dimensions of impact (e.g. effectiveness in achieving attacker goals). These Likert-scale ratings are then aggregated using some (typically nonlinear) function to assign each system feature an overall score. These scores can then be used to rank order

the system features to inform decision makers. Unfortunately, this methodology may fall short of its goal of simplifying the underlying complexity to an actionable set of features.

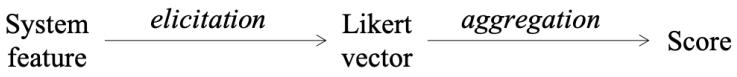


Figure 1: A common approach to system assessment

This approach does not eliminate the complexity introduced by human subjectivity from the process; in fact, no methodology will do so. However, it does "compartmentalize" the subjectivity into various dimensions. This is valuable, because there is likely to be much less variability in the expert rankings along each of these dimensions. It therefore promotes greater consistency of the end results between experts and between repeated assessments by the same expert, giving the decision maker more confidence in the results.

Nevertheless, a reasonable decision maker may retain important skepticism about the results. Significant variability of the final ranking is still possible due to subjective choices made in various aspects of the process. This is often exacerbated by the nonlinearity of the aggregation function. The final results may be more sensitive to uncertainty in some parameters than others. In this document we focus on three important sources of complexity:

1. uncertainty in the Likert ratings assigned to system features,
2. inconsistent interpretations of the relative strength of Likert scores, and
3. subjective choices for the relative weights of each dimension.

The first recognizes that two SMEs may assign different scores to the same system feature, and that a single SME may assign different scores at different times. Therefore, it makes sense not to view the assignment of scores to features as a deterministic function, but rather as a random variable. The sensitivity of the entire evaluation depends, in part, on the distribution of these random variables. We explore methods for eliciting information about these latent distributions, and for evaluating the sensitivity of the final results as we vary the scores in accordance with these distributions.

The second source of variability arises from the fact that, on a 1-5 rating scale, the difference between 1 and 2 may not be the same as the difference between 3 and 4. That is, it may not represent a linear scale. Another way of viewing the same phenomenon is to imagine expanding the scale to values between 1 and 100. A linear scale would translate a rating of 2 into a rating of 25, 3 into 50, 4 into 75, etc. In some settings, it may be clear that 1 and 5 should translate to 1 and 100 respectively, but 2, 3, and 4 may all be mid-range values, mapping, for example, to 40, 50, and 60 respectively. This difference in relative strength of the Likert values can also affect the overall ranking. We suggest methods for accounting for this fact.

Finally, the relative weights of each dimension ranked are often a contentious choice because they codify qualitative assessments using concrete numerical values without much scrutiny regarding their accuracy. By considering these weights explicitly as parameters to the ranking function, we can probe the consequences of re-balancing the weights.

Scope. Some words are in order about a crucial source of skepticism that we explicitly consider out of scope in this document, namely the choice of aggregation function. This includes the decision of what its domain should be, i.e. what dimensions of system features should be rated during elicitation. A common criticism levied against these methodologies is that the process doesn't accurately "get at" the underlying relationships of the real system. That is, regardless of any variability arising from elicitation, the resulting ranking cannot be a faithful representation of the system. While we do address this issue tangentially in our examination of the effect of changing the weights of various dimensions, we explicitly bracket out this larger question of faithfulness for several reasons.

First, and most importantly, our objective here is not to evaluate the particular merits of any given methodology. We take as a given that methodologies that have been repeatedly used across several sponsors provide some value for the decision makers. Rather, our objective is to develop a framework in which *any* similar methodology may support better decisions by providing the decision maker with important contextual information regarding alternative results that could have come out of the same analysis.

Second, there is no single aggregation function that will serve the purposes of all analyses. For example, CJA and ACR have opposite perspectives. The first attempts to get at value from a

defender's point of view, while the latter attempts to get at value from an attacker's point of view. Neither perspective is "correct," and both can provide useful insights into how to prioritize defensive resources. The choice of one over the other will depend on the perspective and priorities of the decision maker.

Finally, we believe that separating out concerns of sensitivity from concerns of faithfulness is useful for conceptual clarity. The former is akin to criticisms about the accuracy of a tool, while the latter is akin to criticisms about the appropriateness of the tool. Truly answering criticisms about appropriateness may require philosophical discussions around the theory of measurement. Addressing the accuracy of a tool is a much easier problem. For example, even with a grossly inaccurate ruler, one can tell if a person is growing.

Example Methodology

Although the ideas in this document are designed to apply to a wide range of methodologies incorporating the same core process, it is instructive to have a concrete example with which to work through the ideas. To that end, we provide some basic background regarding Adversary-Driven Cyber Resiliency (ACR) (McQuaid et al., 2017).

The ACR methodology is a structured means of identifying which attacks against a target system are most likely to be pursued by an advanced adversary. While an important and challenging part of the process is identifying a set of attacks against a target system, here we start with the assumption that such a set has already been identified. This allows us to focus on the means by which this set is put into a rank-ordered list with the aim of identifying the highest priority attacks.

Once the set of attacks $A = \{a_1, \dots, a_n\}$ has been identified, evaluators elicit from subject matter experts a 1-5 (low-to-high) ranking of each attack according to two categories: difficulty and impact. These two categories are further subdivided into sub-categories. The overall difficulty is based on the difficulty in each of five different dimensions: knowledge of vector, development skill, access or installation, survivability or detectability, and command and control. The overall impact is similarly based on the impact in each of three different dimensions: scope, mission impact, and defender impact.

For our present purposes, the exact meaning of each of these eight dimensions is not important. Rather, we note that the elicitation process implicitly defines a function

$$\boldsymbol{e}: \boldsymbol{A} \rightarrow \boldsymbol{D^5} \times \boldsymbol{I^3} \tag{1}$$

Where $D = I = \{1,2,3,4,5\}$. This represents the first arrow of Figure 1 which maps a system feature (in this case a possible attack) to a vector $(d_1, d_2, d_3, d_4, d_5, i_1, i_2, i_3) \in D^5 \times I^3$ of Likert-scale values.

The remainder of the ACR methodology consists of computing a single value—the return on investment, or ROI—for each attack based on the Likert vector produced by e. The ROI is defined as follows.

$$\boldsymbol{ROI}(\boldsymbol{d_1}, \boldsymbol{d_2}, \boldsymbol{d_3}, \boldsymbol{d_4}, \boldsymbol{d_5}, \boldsymbol{i_1}, \boldsymbol{i_2}, \boldsymbol{i_3}) = \frac{\boldsymbol{i_1}+\boldsymbol{i_2}+\boldsymbol{i_3}}{\boldsymbol{d_1}+\boldsymbol{d_2}+\boldsymbol{d_3}+\boldsymbol{d_4}+\boldsymbol{d_5}} = \frac{\sum_{j=1}^{3} \boldsymbol{i_j}}{\sum_{k=1}^{5} \boldsymbol{d_k}} \tag{2}$$

Using the ROI, each attack is assigned a single value which is then used to rank the attacks. Figure 2 shows a row from a matrix of attacks annotated with the SME assessments and the resulting ROI. The rows of such a matrix can be sorted, allowing those attacks with the highest ROI to bubble to the top. Key stakeholders can use this prioritized list of attacks to help them make decisions regarding how and where to spend limited funds on cyber resiliency solutions.

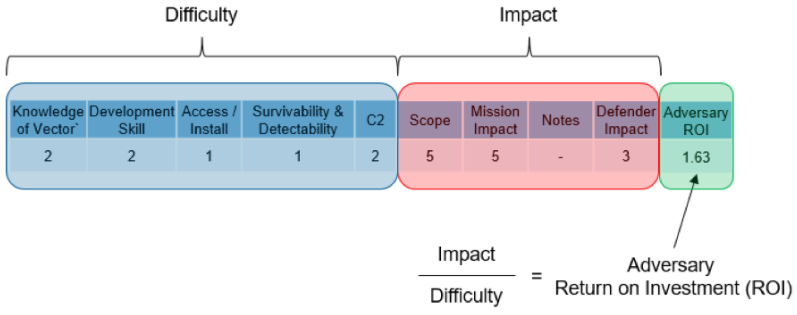


Figure 2: Sample row from matrix of identified attacks.

From concrete to general. We choose to present ACR as an introductory example due to its relative simplicity. Many other methodologies incorporate aspects that have the same basic structure. There is an elicitation phase that implicitly defines a function e that assigns a Likert vector to some set of system attributes. This is followed by an aggregation phase that applies

some aggregation function f to the output of e creating a rank-ordered list of the original set. Thus, the whole process of assigning an individual attribute a numerical value is summarized as

$$\boldsymbol{a} \mapsto \boldsymbol{f}\big(\boldsymbol{e}(\boldsymbol{a})\big) \tag{3}$$

The rank order is simply the result of sorting the outputs $f\big(e(a)\big)$.

The variability in the outcome can be traced back to several ways in which human subjectivity is encoded in Eq. $\boldsymbol{a}\mapsto\boldsymbol{f}\big(\boldsymbol{e}(\boldsymbol{a})\big)$ (**3**). Firstly, e encodes a good amount of the subjectivity of the SMEs. The above explanation treats e as a deterministic function, when it would be more appropriate to think of it as a randomized function. That is, we should view $e(a)$ as a random variable. Since $e(a)$ is actually a vector, we might think of each component as being a random variable (possibly, though not necessarily independent from each other component).

The other sources of subjectivity are implicitly hidden in f, which encodes the relative differences between the Likert-scale values, as well as the relative weights of each of the rating dimensions. That is, there are hidden parameters not shown in Eq. $\boldsymbol{a}\mapsto\boldsymbol{f}\big(\boldsymbol{e}(\boldsymbol{a})\big)$ (**3**), so that it might be more informative to write it as

$$\boldsymbol{a} \mapsto \boldsymbol{f}(\boldsymbol{e}(\boldsymbol{a}), \overline{\boldsymbol{w}}, \boldsymbol{g}) \tag{4}$$

where $\overline{w}$ is a vector of multiplicative weights (one for each dimension) and g is a function that re-scales the Likert values.

Using this more general form, the ACR evaluation function could be written as

$$\boldsymbol{ROI}(\boldsymbol{d_1}, \boldsymbol{d_2}, \boldsymbol{d_3}, \boldsymbol{d_4}, \boldsymbol{d_5}, \boldsymbol{i_1}, \boldsymbol{i_2}, \boldsymbol{i_3}) = \frac{\boldsymbol{i_1}+\boldsymbol{i_2}+\boldsymbol{i_3}}{\boldsymbol{d_1}+\boldsymbol{d_2}+\boldsymbol{d_3}+\boldsymbol{d_4}+\boldsymbol{d_5}} = \frac{\sum_{j=1}^{3} \boldsymbol{w_i}\cdot\boldsymbol{g}(\boldsymbol{i_j})}{\sum_{k=1}^{5} \boldsymbol{w_k}\cdot\boldsymbol{g}(\boldsymbol{d_k})} \tag{5}$$

where each i_j and d_k are random variables with some underlying distribution.

Assessor Uncertainty

Since we are driving our assessments with SME input, and we are cognizant of the complexity of that sampling methodology, we must develop a way of extracting information which is congruent with our understanding. To that end, we will focus on a simple approach to capturing the complexity of psychological uncertainty which the decision-maker is experiencing during the

information elicitation process. We propose a modification to the elicitation process that allows us to transform the function *e* above from being deterministic to being probabilistic. When applying complexity science, agency is often modeled using probabilistic decision spaces (Norman et al., 2018), so applying a probabilistic paradigm is appropriate here.

1. In addition to the original assessment criteria, ask the assessors to categorize their certainty of each response:

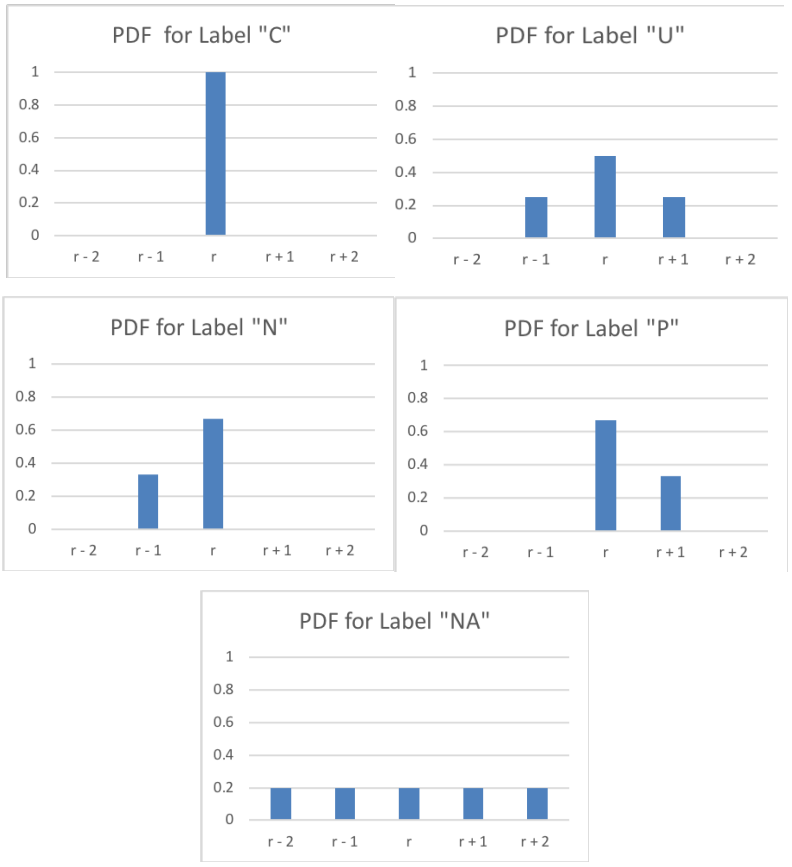


Figure 3: Notional label probability density functions (PDFs)

   a. Certain (C)—There is no ambiguity in the response; the assessor is 100% confident in the response.
   b. Uncertain (U)—Assessor is uncertain about the response. The label indicates the response value may be the indicated response ± 1.
   c. Positive Skew (P)—This label indicates that the response value may be the indicated response, or the indicated response +1.

d. Negative Skew (N)—This label indicates that the response value may be the indicated response, or the indicated response –1.
e. No Opinion (NA)—This response is reserved for cases where the assessor has no knowledge of the proper response.

2. Model each label as a discrete probability distribution. Figure 3 depicts notional probability density functions (PDFs) that can be associated with each level of uncertainty.

Table 1: Attack assessment example

Dimension	Score	Uncertainty Label	
Knowledge of vector	2	N	Difficulty
Development Skill	2	U	
Access	1	C	
Survivability	1	C	
Command & Control	2	N	
Scope	5	N	Impact
Mission Impact	5	N	
Defender Impact	3	U	

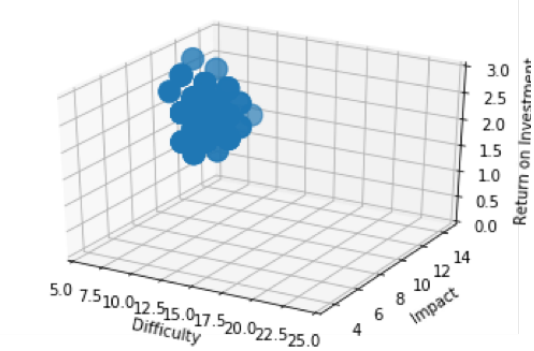
3. Sample from these distributions to generate a statistical ensemble representative of the attack under assessment. The attacks can then be rank-ordered according to their average ROI. Alternatively, an ensemble of rank orderings can be created by repeatedly sampling once for each attack and rank ordering the results.

Note that we consider only discrete probability distributions since the response scale is not continuous. The observations generated by the sampling process can be used as a proxy to

characterize assessor indecision or uncertainty. In this case, higher variance in the outcomes represents higher degrees of uncertainty on the part of the assessor.

Sampling from a single row. Consider an example ACR assessment looking at a specific type of SYN flooding attack against a server. The assessor goes through the columns of the ACR matrix assigning values from 1 to 5 and applying an uncertainty label to each. Table 1 shows the assessment scores and labels assigned for this example.

For this example, the ROI score provided by ACR is 1.625. The addition of the labeling process introduces probability distributions for each of the values, allowing an assessor to sample from each space many times. Such an ensemble of samples provides additional



(a)

ROI (Calculated)	1.625
μ^2_{ROI}	1.704
σ^2_{ROI}	0.071
$Max(ROI)$	2.600
$Min(ROI)$	1.222
$\sigma^2_{Difficulty}$	0.914
σ^2_{Impact}	1.020

(b)

Figure 4: Simulations results (a) and statistics (b)

assessment metrics, as well as a means for assessors to more easily visualize alternative outcomes. Figure 4(a) shows the observed outcomes of the sampling process, while Figure 4(b) shows some of the resulting statistics. These results reflect the assessor's sentiment provided in the labeling; the mean ROI is noticeably higher than the calculated ROI, likely due to the negative-leaning difficulty labels. Additionally, the difficulty and impact variances are relatively high, reflecting a degree of uncertainty or indecision on the part of the assessor.

In contrast to this example, we could perform the same exercise with labels reflecting greater certainty on the part of the assessor. The new labels are shown in Table 2, and the results are shown in Figure 5. The statistics in Figure 5(b) show a mean closer to the calculated ROI

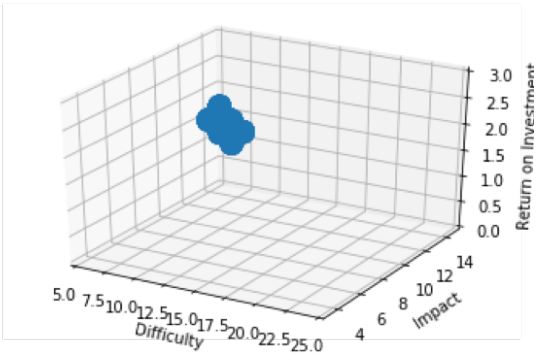
(though lower because of the negative skew on the impact column), and a much lower variance in both difficulty and impact.

Alternative rankings. The example above demonstrates how one can use the certainty labels provided by SMEs during the elicitation process to assess the uncertainty associated with the ROI of any given attack. Generally speaking, a greater range between the max and min values and a greater variance indicate less certainty in the ROI value calculated from the assessment scores. Ultimately, however, a system owner will have to decide which attacks to prioritize when considering possible mitigations. Therefore, it is important to generate alternative prioritizations that might result by sampling from the distributions associated with each row.

Table 2: Attack assessment example: Increased certainty

Dimension	Score	Uncertainty Label	
Knowledge of vector	2	C	Difficulty
Development Skill	2	C	
Access	1	C	
Survivability	1	C	
Command & Control	2	U	
Scope	5	N	Impact
Mission Impact	5	C	
Defender Impact	3	C	

One straightforward method to generate an alternative ranking would be to generate n samples from each row, calculate the average ROI, and order the rows according to these average results. While this approach has some intuitive appeal, it has some drawbacks. Notably, much of the information contained in the range and variance of outcomes for any row is lost by averaging. It also generates a single alternative ranking to the one determined by the calculated ROIs rather than exhibiting the wider range of possible outcomes. At issue is the fact that



(a)

ROI (Calculated)	1.625
μ^2_{ROI}	1.588
σ^2_{ROI}	0.023
$Max(ROI)$	1.857
$Min(ROI)$	1.133
$\sigma^2_{Difficulty}$	0.502
σ^2_{Impact}	0.212

(b)

Figure 5: Increased certainty: Simulation results (a) and statistics (b)

ranking averages of alternative outcomes is typically not the same as averaging the ranking of alternative outcomes, especially when the outcomes are determined by nonlinear functions.

We therefore propose that an ensemble of alternative rankings be generated in the following manner. Generate one sample from each row to determine a possible ROI value for every row. Rank the rows according to these samples. Repeat the above process n times to generate n alternative rankings. These rankings will typically exhibit more of the possible variety of outcomes, giving the assessors a better idea of which attacks remain near the top of the rankings in all or most of the outcomes, and which ones are less certain to rise to the top. One could evaluate these alternatives qualitatively, or one could additionally average the rankings to generate an average summary of the possible outcomes.

Model Uncertainty

The focus of this section is on capturing the effect on decision makers due to uncertainty in the fixed model parameters w_i and g.

Recall that w_i is a relative weight of input parameter i. For example, in the context of ACR, if a decision maker judges that an adversary's ability to survive on the system is twice as important as the ability to gain access in the first place, then $w_{survivability}$ would be $2 \cdot w_{access}$. ACR sets all the default weights to 1.

Similarly, g is a scaling function that adjusts the relative weight of each Likert rating independent of to which parameter it applies. By default, ACR sets g to be the identity function. However, by replacing g with a nonlinear function such as the logistic function depicted in Figure 6, the jump from a 1 to a 2 could be much less impactful than the jump from a 3 to a 4.

These model parameters are typically fixed in advance by the assessment methodology. As a general rule, changes to such fixed parameters should not be undertaken lightly. Nevertheless, it has been our experience that a major source of skepticism regarding the applicability of an assessment methodology pertains to the degree to which these model parameters reflect underlying relationships about the system under study. For that reason, it is valuable to assess the sensitivity of the final recommendations to changes in these values.

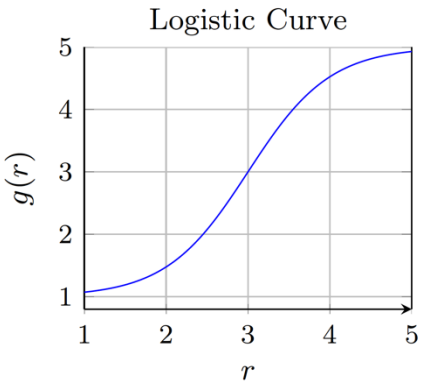


Figure 6: Rescaling of 1-5 ratings r **by logistic function** g

Assuming decision makers buy into the general approach of a methodology, demonstrating that reasonable adjustments to these parameters have little effect on the outcomes would go a long way toward increasing confidence in decisions.

The adjusting of model parameters typically interacts with assessor uncertainty in important ways. For example, if survivability is weighted very heavily, then a small amount of uncertainty in that parameter could generate outcomes with a much bigger variance than if it had a small weight. Likewise, when the jump from a rating of 3 to 4 is bigger than the jump from 1 to 2, uncertainty surrounding items rated as 1 will have less impact than uncertainty surrounding items rated as 3. This means that the relative sensitivity to changes in model parameters is not an

inherent feature of any methodology. We therefore propose that assessors explore the effects of changing parameters once the elicitation process (with certainty labels) has occurred.

Mechanically, an assessor would perform the process described in the previous section for each selection of model parameters of interest. The results can then be qualitatively compared. Given the assessors inputs, do the same attacks rise to the top of the prioritization? Are there some attacks whose ranking is stable under one set of parameters, but uncertain under a different choice? If there is great variability in the overall recommendations, this should be taken into account when deciding which attacks to focus on first. Alternatively, if the overall recommendations are relatively unchanged regardless of model parameters, this can provide a greater level of certainty regarding those recommendations.

Alternative Approaches

The previous two sections detailed particular recommendations that would be easy to incorporate into existing cyber assessment methodologies with limited effort and disruption. The association of probability distributions with SME inputs allows one to generate alternative recommendations and qualitatively assess the sensitivity of the results to changes in the inputs.

One reasonable objection to our approach would be to note that there is considerable skepticism in the social sciences regarding the value of quantitative statistics applied to ordinal (Likert-scale) data. Since ordinal data is not continuous, calculating the means and variances could lead to misleading or meaningless values (Mu et al., 2012). For this reason, we briefly mention some alternative approaches to addressing the limitations inherent in ordinal data.

Factor analysis. Factor analysis can be used to take observed variables and represent them in terms of their underlying latent factors. Factor Analysis (FA) is a dimensionality reduction technique that yields apparently similar results to Principle Component Analysis (PCA). The two techniques are often applied interchangeably in social sciences, though there is some dispute as to whether PCA is appropriate for all applications (van der Eijk and Rose, 2015). Lawley and Maxwell present a detailed mathematical background for FA in their paper (1962).

Software factor analysis packages are readily available; Scikit Learn for Python contains several FA models, and IBM's SPSS® statistical software package is available commercially. From a practical standpoint, this means that an analyst only needs to provide a correlation matrix of the input variables to perform factor analysis (Costello and Osbourne, 2005; Flora et al., 2012).

Reliability. Statistical reliability metrics such as Cronbach's α are often used in order to evaluate internal consistency of test measures in the fields of psychology, sociology, and medicine (Tavakol and Dennick 2011). Cronbach's α is only valid for continuous data, but similar analogs exist that may be used for the ordinal data generated by assessments such as ACR. Zumbo et al. propose a method for calculating ordinal reliability for a score of p items:

$$\boldsymbol{\alpha} = \frac{p}{p-1}\left[\frac{p(\bar{f})^2 - \bar{f}^2}{p(\bar{f})^2 + \bar{u}^2}\right] \tag{6}$$

where $\bar{f}$ is the average of p factor loadings and $\bar{u}$ is the average of p uniqueness (Zumbo et al., 2007).

Ordinal α could be used to quantify assessor agreement in cases where multiple assessors are evaluating the same system, where a single assessor evaluates the same system multiple times, or in the presence of a mature dataset of completed evaluations, potentially to compare evaluations against similar systems.

Robust optimization. Relying solely on the possible range of ratings assigned to values, one can avoid performing statistical analysis over the ordinal data completely and focus on optimizing one's decisions in light of the worst-case errors in some bounded number of input values. More explicitly, if a range of values is associated with each input rating, it is possible to identify the largest deviation from the default ranking assuming n of the inputs might be at the extreme end of their range. Decision makers may opt to prioritize their focus based on an estimate of how bad each investment is in light of the worst-case combination of inputs at their extreme values (Gorissen et al., 2015).

Computational social science and agent-based modeling. Agent-based modeling (ABM) is beginning to show promise as a tool for computational social scientists to begin to explore the

impact of social humans on the creation of resilient cyber systems by shedding light on the impact of human decision-making on the system, as well as the impact that various policies will have on the human decision-makers (Norman and Koehler, 2017). A similar approach, in which the assessments being used are augmented with an ABM of the human touchpoints in the cyber system, could be applied here. This ABM could be informed by the same SME-driven data collection methodology as ACR.

By assessing the cyber resiliency of a system from the perspective of complexity science we are encouraged to begin to incorporate the probabilistic nature of agency, which is a hallmark of complexity and the foundation of all complex adaptive systems (Norman et al., 2018).

Conclusion

ACR-style assessments are attractive in that they do not require a lengthy elicitation process (as is the case with Crown Jewels Analysis), but the lack of such a process means that other strategies must be considered to ensure that cyber-assessments are providing accurate, appropriate, and valuable information. In this chapter, we have investigated several methods for improving cyber resiliency assessments relying on ordinal, Likert-scale expert responses. Though ACR formed the basis for the examples in this paper, these examples could be modified or extended to encompass other Likert or ordinal response methodologies. The brief overview of statistical techniques aimed at analyzing Likert scale data is just a small view of available techniques. Computational and traditional social sciences, as well as psychological and medical research, rely heavily on this type of data, and there is a wealth of science and technology that could be utilized to make these types of assessments more robust.

Disclaimer

The authors' affiliation with The MITRE Corporation is provided for identification purposes only, and is not intended to convey or imply MITRE's concurrence with, or support for, the positions, opinions, or viewpoints expressed by the authors. This chapter has been approved for Public Release; Distribution Unlimited; Case Number 18-2522-2.

References

Costello, Anna B., and Jason W. Osbourne. 2005. "Best practices in exploratory factor analysis: Four recommendations for getting the most from your analysis." *Practical Assessment, Research & Evaluation* 10 (7).

Flora, David B., Cathy LaBrish, and R. Philip Chalmers. 2012. "Old and new ideas for data screening and assumption testing for exloratory and confirmatory factor analysis." *Frontiers in Psychology* 3 (55).

Gorissen, Bram L., Ihsan Yanikoglu, and Dick den Hertog. 2015. "A practical guide to robust optimization." *Omega* 53: 124-137.

Lawley, D. N., and A. E. Maxwell. 1962. "Factor analysis as a statistical method." *Journal of the Royal Statistical Society. Series D (The Statistician)* 12 (3): 209-229.

Liebovitch, Larry S., Paul R. Peluso, Michael D. Norman, Jessica Su, and John M. Gottman. 2011. "Mathematical model of the dynamics of psychotherapy." *Cognitive Neurodynamics* 5 (3): 265-275.

McQuaid, Rosalie M., Jeffrey Picciotto, Thomas E. Everett, Duane A. Souder, Paul D. Rowe, Sharon Orser-Jackson, and Daniel E. Fitzpatrick. 2017. *Adversary-driven cyber resilience (MP 170687).* Technical report, The MITRE Corporation.

Mu, Mu, Andreas Mauthe, Gareth Tyson, and Eduardo Cerqueira. 2012. "Statistical analysis of ordinal user opinion scores." *IEEE Consumer Communications and Networking Conference (CCNC).* 331-336.

Norman, Michael D., and Matthew T.K. Koehler. 2017. *Cyber defense as a complex adaptive system: A model-based approach to strategic policy design.* Technical report, arXiv.

Norman, Michael D., Matthew T.K. Koehler, and Robert Pitsko. 2018. "Applied complexity science: Enabling emergence through heuristics and simulations." *Emergent Behavior in Complex Systems Engineering: A Modeling and Simulation Approach* (Wiley).

Tavakol, Mohsen, and Reg Dennick. 2011. "Making sense of Cronbach's alpha." *International Journal of Medical Education* 2: 53-55.

van der Eijk, Cees, and Jonathan Rose. 2015. "Risky business: Factor analysis of survey data assessing the probability of incorrect dimensionalisation." *PLOS ONE* 10.

Watters, Chalton J., and Shaun P. Morrissey. 2018. *Crown Jewels Analysis (CJA) Process (MP 180005).* Technical report, The MITRE Corporation.

Wynn, Jackson, Joseph Whitmore, Geoff Upton, Lindsay Spriggs, Dan McKinnon, Rich McInnes, Rich Graubart, and Lauren Clausen. 2011. *Threat Assessment and Remediation Analysis (TARA) (MTR 110176).* Technical report, The MITRE Corporation.

Zumbo, Bruno D., Anne M. Gadermann, and Cornelia Zeisser. 2007. "Ordinal versions of coefficients alpha and theta for Likert rating scales." *Journal of Modern Applied Statistical Methods* 6 (1).