Problems with Risk Matrices Using Ordinal Scales

Michael Krisper
Institute of Technical Informatics
Graz University of Technology
Graz, Austria
michael.krisper@tugraz.at

***Abstract*—In this paper, we discuss various problems in the usage and definition of risk matrices. We give an overview of the general process of risk assessment with risk matrices and ordinal scales. Furthermore, we explain the fallacies in each phase of this process and give hints on which decisions may lead to more problems than others and how to avoid them. Among those 24 discussed problems are ordinal scales, semi-quantitative arithmetics, range compression, risk inversion, ambiguity, and neglection of uncertainty. Finally, we make a case for avoiding risk matrices altogether and instead propose using fully quantitative risk assessment methods.**

***Index Terms*—Risk Matrix, Risk Assessment, Risk Metric, Ordinal Scales, Range Compression, Consistency, Quantitative Methods, Qualitative Methods, Semi-Quantitative Methods, Human Bias**

I. Introduction

Risk matrices are established tools to assess and rank risks in many domains and industries. They have become so common that everyone accepts and uses them without question. They have many seemingly benefits like the simplicity of usage, different coloring systems with traffic light semantics, intuitive understanding, and are seemingly proven in use over many decades. When there is little or no data available, they are praised as the weapon of choice for tackling risks and estimates in projects. Nevertheless, they have many flaws and problems that will be covered in this work.

More than a decade ago, Anthony Cox and his team started a riot against risk matrices which has not come to an end since [1]–[5]. He has shown and proven that risk matrices have severe problems that could diminish their usefulness to the point where they become even worse than random. More and more scientists, engineers, and managers have since supported the cause against risk matrices, and amongst them, Douglas Hubbard is one of the most prominent ones. In his excellent book series "How to measure anything" [6], [7], he also defends Cox et al. and demonstrates some ideas and techniques for a quantitative risk assessment method to overcome and avoid the problems of classical risk matrices.

In this work we build upon the findings of Cox [3], Hubbard [7], Artzner [8], Talbot [9], Kahnemann & Tversky [10], as well as many others over the last few decades. We show that risk matrices today still have some flaws, fallacies, and pitfalls and explain what those are. By showing them, we want to, once more, state a case for fully quantitative risk assessment using quantitative value ranges, ratio scales, and probability distributions, which are considering the uncertainties throughout the risk analysis. Our focus in this paper lies in summarizing pitfalls and fallacies in risk assessment using ordinal scales and risk matrices with concise and understandable explanations and examples.

The paper is structured as follows: After this introduction, the motivation sections show some examples of what can go wrong using risk matrices and its consequence. Subsequently, we directly dive into the overview and descriptions of the problems, pitfalls, and fallacies of risk matrices found in the literature.

The authors are researchers at the Graz University of Technology and have a background in automotive safety, quality, and security. This work aims to show the problems of qualitative risk assessment methods to argue towards quantitative methods. In particular, in our research group, we are currently working on a method for integrated quantitative risk assessment [11], which combines safety and security. For that, we are developing a tool based on attack-trees using truly quantitative methods called RISKEE [12].

II. Motivation

"What is so bad about risk matrices?", one may ask, *"they are so widely accepted and established tools, they cannot be wrong."*. Only because something is established does not mean it is without any flaws. In this section, we show some examples of pitfalls that may occur when using risk matrices. We will show some artificial examples and real-life use-cases, where the ranking of risks with risk matrices is illogical, unreasonable, or leads to problems.

A. Oil Leakage and the MIL-STD882C

This example was taken from *"What's wrong with risk matrices"* by Cox et al. [3] and shows the problem of risk inversion (see the problem (U)). In this example, two physical hazards for environmental damage (fuel leakage in this case) are compared. The first event consists of 1 ounce of fuel spills five times per hour. The second event causes more damage but happens less frequently, with 10 pounds of fuel leaking once per week. According to the military standard, 882C [13], both would arguably get the highest frequency rating, but the one leaking 1 ounce would get a negligible hazard rating (resulting in a MEDIUM score). In contrast, the 10-pound event would get a marginal or even critical severity (resulting in a HIGH score). If we compute the risks quantitatively, we get another

HAZARD CATEGORY / FREQUENCY	CATASTROPHIC	CRITICAL	MARGINAL	NEGLIGIBLE
FREQUENT	HIGH	HIGH	HIGH	MEDIUM
PROBABLE	HIGH	HIGH	MEDIUM	LOW
OCCASIONAL	HIGH	HIGH	MEDIUM	LOW
REMOTE	HIGH	MEDIUM	LOW	LOW
IMPROBABLE	MEDIUM	LOW	LOW	LOW

Fig. 1. An example of a risk matrix defined in the standard MIL-STD-882C [13].

result: The 1-ounce event produces 52.5 pounds of leakage per week (1oz*5*24*7), while the 10-pound event leaks 10 pounds. Thus, the first event should be rated way higher than the second event, which it is not. Figure 1 shows the risk matrix taken from MIL-STD-882C. This example shows cases where the qualitative risk score does not reflect the actual quantitative risk. Even worse, it results in an inverse order for the events' priorities, which is the consequence of risk inversion.

B. Failure Mode and Effects Analysis (FMEA)

Another example is the risk matrix in the Failure Mode and Effects Analysis (FMEA) [14]. Especially here, the severity scale is problematic because it combines four different effects scales into one ordinal scale and assigns them ranks from 1 to 10 (annoyances, failure of secondary functions, failure of primary functions, failure of safety). Furthermore, this ordinal scale is multiplied with other influence factors to get a resulting risk priority number (RPN), although multiplication is not defined on ordinal scales. Furthermore, a higher detectability factor could reduce the RPN tremendously but does not reduce the actual hazard. Is the risk of a hazard less severe just because we can detect it?

C. Hazard and Risk Analysis (HARA)

The Hazard and Risk Analysis (HARA) [15] is done at the concept and system level in the early stages of product development. The problem of this method is the ambiguity of the input scales, in particular, exposure. First, let us summarize the method itself: During the analysis, one assesses possible hazardous scenarios for their severity, exposure, and controllability. All three values are logarithmically distributed ordinal scales that assign a number for the rank. The ranks for the individual scales get added up, and the resulting number is translated into an ASIL (automotive safety integrity level) classification. Depending on the ASIL, there are exponentially more complex requirements to fulfill for developing a product in a safe way. These requirements become so high that it is challenging to implement them using only a single component. Thus, the ASIL can be decomposed into subsystems with lower ASIL but have to be independent, redundant, and diverse to avoid common cause failures. This decomposition increases the costs tremendously. A false ranking of the initial values has severe consequences to all subsequent development efforts

Class	E1	E2	E3	E4
Description	Very low probability	Low probability	Medium probability	High probability
Definition of frequency	Situations that occur **less often than once a year** for the great majority of drivers	Situations that occur **a few times a year** for the great majority of drivers	Situations that occur **once a month or more often** for an average driver	All situations that occur during **almost every drive on average**

Fig. 2. Informative examples for the exposure in the ISO 26262 [15].

C1	S1	S2	S3
E1	QM	QM	QM
E2	QM	QM	QM
E3	QM	QM	A
E4	QM	A	B

C2	S1	S2	S3
E1	QM	QM	QM
E2	QM	QM	A
E3	QM	A	B
E4	A	B	C

C3	S1	S2	S3
E1	QM	QM	A
E2	QM	A	B
E3	A	B	C
E4	B	C	D

Fig. 3. The risk matrices for the Hazard and Risk analysis in ISO 26262 [15], illustrated here by splitting it up into three parts with different controllability scores.

and costs of a product. Especially border cases are the problem here: Decreasing the score of a borderline case could decrease the resulting ASIL from D to C, which cuts the effort for product development to half. Let us examine this in the case of the exposure score. The exposure can be defined in two ways: either via the frequency of occurrence over time or as the proportion of duration in hazard situations compared to the total operating time of a product. These two aspects are reasonable because sometimes the frequency is needed(e.g., traffic situations), and sometimes the event's duration (e.g., radiation exposures). Here, ambiguity strikes the hardest: Changing the argumentation from one to the other could change the score entirely. Furthermore, even when staying in the same category, the scales are ambiguous. Figure 2 shows an excerpt for the exposure from the informative annex of the ISO 26262 [15]. Exposure rank 2 states *a few times per year for most drivers*, while E3 states *once a month or more often for an average driver*. There are two ambiguities here: Firstly, what exactly is the definition of the majority of drivers and the average driver? Does the *majority of drivers* mean more than 80%, 90%, 99%, or is it 51%? Secondly, where is the border between a few times per year and once a month or more often? Is it six times per year? Even if the boundaries would have been defined exactly, a quantification problem is still left, which we will discuss later on (problem (J)).

III. The Problems of Risk Matrices

In the following sections, we go over the typical process of risk scoring methods based on risk matrices and explain why this is problematic and what the problems are. Furthermore, we compare this to quantitative risk assessment methods to show that they do not suffer from the described problems and should always be preferred over qualitative methods.

Qualitative (or semi-quantitative) risk assessment methods based on ordinal scales and risk matrices typically are done in five phases, which is illustrated by Figure 4. It shows an overview of the five phases and enlists the problems that may occur in each phase. Just to give a comparison, Figure 5 shows the corresponding quantitative approach, which also has five phases. The risk score is computed quantitatively by

estimating plausible ranges of input factors, simulating them using Monte-Carlo simulation, and comparing them to the risk appetite (also called risk affinity) using a loss exceedance curve [7], [12].

Nevertheless, in the following sections, we will discuss the five phases of qualitative risk assessment and their problems and compare them to the quantitative approach whenever reasonable:

- ***Phase 1: Identifying the influence factors***. First, a set of influential factors has to be defined.
- ***Phase 2: Rating of the factors***. In this phase, the input values are rated according to some scale.
- ***Phase 3: Combining the ratings***. The scorings are combined to get a final risk score.
- ***Phase 4: Ranking the combinations***. The risk score is ranked against other risks or filtered by some threshold.
- ***Phase 5: Decision making based on the rankings***. In the end, decisions have to be made which risks to reduce. The goal is to bring the risks down to a tolerable level.

Phase 1: Identifying the influence factors

Before doing any risk assessment, one has to define the influence factors that affect the system's risk, which is assessed. Most standardized risk assessment methods defined the influence factors right away, e.g., impact, probability, severity, exposure, utility, loss. Some leave it open to be defined by the practitioners. Others are only defined within a single organization to be specialized for a specific situation, e.g., for actuary sciences, insurances, medicine, or the financial sector. For such industries, the identification of influence factors plays a massive role in risk assessment. The number of influence factors often is related to the used method. General risk assessment methods tend to use only two or three factors; specialized ones tend to use more. Furthermore, multiplicative methods also tend to use lesser factors, like two or three, and additive ones use more in general. Examples of the most frequently used factors for general risk assessment methods are the following:

- *Impact*: The impact corresponds to the actual outcome when the risk event occurs. A higher impact also means higher risk. This is also called severity, loss, magnitude, harm, effect, threat, consequence, or utility.
- *Probability*: The probability defines some notion of the likelihood that an event occurs. A higher probability also means higher risk. This is often also called exposure, likelihood, vulnerability, frequency
- *Control*: The control factor corresponds to a risk reduction possibility. Higher control over a situation results in lowering risk. Often this is also called controllability, mitigation, reduction, protection, detection, or reaction. Many methods omit this factor by reasoning that its behavior can also be modeled by reducing the probability or impact.

In contrast to that, many additive methods tend to use more specific factors like demographic features, medical attributes, lifestyle attributes. A recently highly discussed example in Austria was the introduction of a rating scheme for the unemployment office in 2018 (Arbeitsmarktservice, AMS) [16]. This was a weighted additive scheme for rating the risk of future unemployment (or otherwise put: the chance for employability). It was based on several demographic and social factors, including, e.g., gender, age, disability, career, education, and many more. The weights for combining the factors were inferred via historical data and statistics and reflected society's bias explicitly. For example, in that scoring algorithm, women have lower job chances than men. This resulted in public discussions, similar to that for amazon's automatic firing algorithm [17]. More about that subject will be discussed in phase 3: *Combining the ratings*.

Now we discuss the problems which may occur in this first phase of selecting the influence factors:

(A) Incompleteness: To do a useful risk assessment, all essential factors have to be accounted for in the analysis. However, sometimes, factors are forgotten or overlooked. This could happen unknowingly or due to ignorance or inexperience. It could be complicated to determine the significant factors that influence the risk, especially for complex behavioral or technical systems. This is not only a problem for qualitative approaches but may happen in quantitative approaches also.

(B) Correlations: The selected factors could be correlated to each other, or in other words: they could influence each other. Even more dangerous is a negative correlation: If one factor grows, others may decline. This could result in worse than random results [3]. There is also a common pitfall of correlated influence factors regarding the view of information theory: When they are strongly correlated, they do not deliver more information than one of them alone. If two scales would always show the same value, it would suffice to use just one because the information gain would be the same. So the solution to this would be to try to avoid correlated factors. This is hard to detect for qualitative approaches, but for quantitative approaches, this could be detected via statistical methods (see problem P for more information).

(C) Irrelevance: Irrelevant factors make the risk assessment more difficult because they have to judge, evaluated, and discussed but have no real impact on the result. There are two aspects of this: The first aspect is real irrelevance - a factor does not increase or decrease the resulting risk. Then it can be skipped in the analysis. The second aspect is, if a factor is the same for all risks, it has no relevance anymore. For example, if we would use a priority factor for risks, and every risk would be rated as a high priority, this does not help the final ranking because every risk would be equal.

(D) Nonlinear Behavior: An input factor could have non-linear behavior, making it difficult to model or rate in the next phase. Logarithmic or polynomial scales could cope with this. However, the more complex a factor is, the more difficult it is to model and judge, e.g., the driving speed as a risk factor is perceived as linear. However, the actual risk increases quadratically or even exponentially [18]–[20].

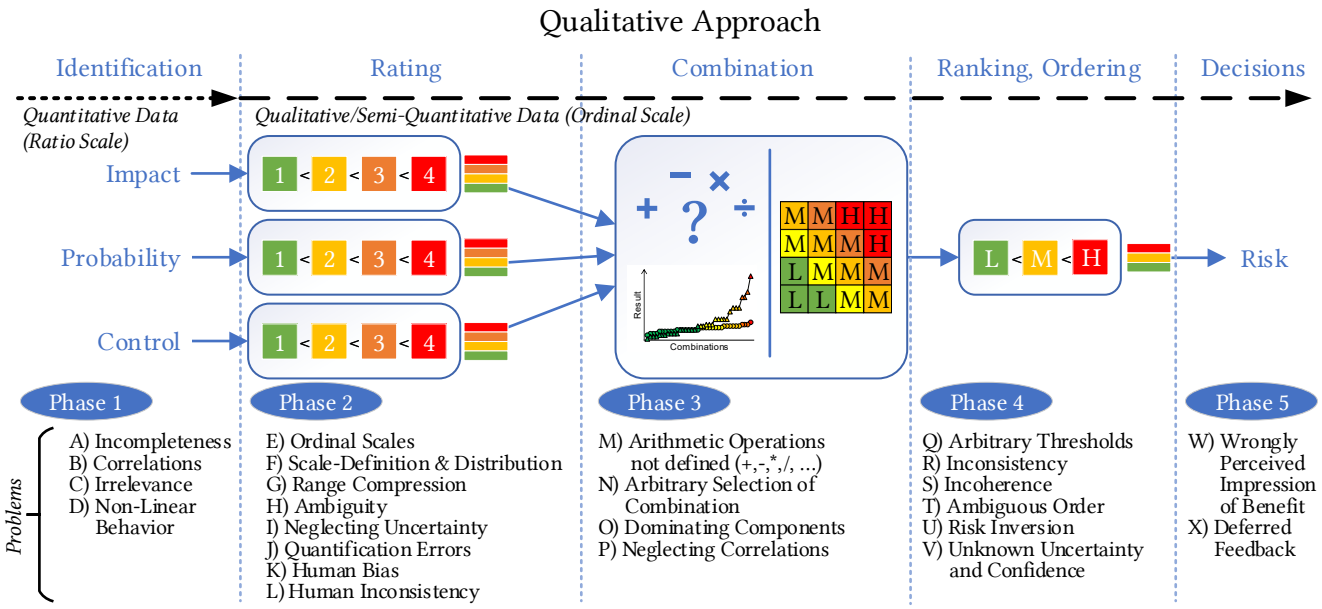

Fig. 4. The flow of information and the processing phases during a typical qualitative risk asessement approach based on ordinal scales and risk matrices.

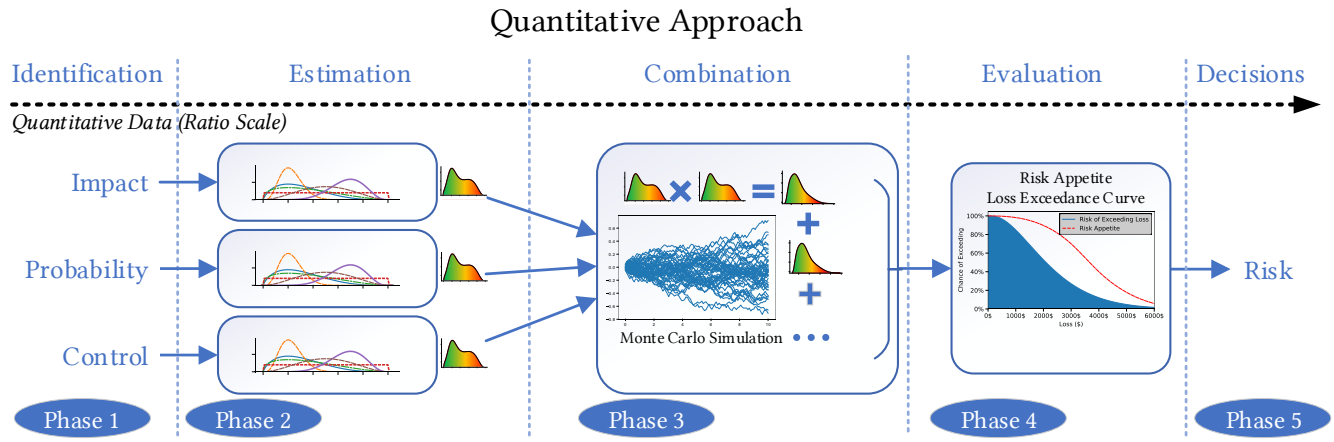

Fig. 5. The flow of information in a quantitative risk assessment approach using ratio scales, probability distributions, monte carlo simulation, and loss exceedance curves.

The mentioned problems apply to qualitative and quantitative risk assessment methods. Nevertheless, quantitative methods can at least tackle them by using mathematical tools. With statistical sensitivity analysis, correlations and irrelevances can be detected (ANOVA, Correlation Coefficients, Hypothesis Tests). Furthermore, non-linear behavior can be modeled in mathematical equations in quantitative models, which would be difficult in qualitative ordinal scales. Only incompleteness is a problem that is hard to solve for both methods. It is not trivial to detect that a factor is missing, which boils down to the often-cited management mantra by Tom DeMarco: *"You cannot control what you cannot measure"* [21]. Nevertheless, Hubbard et al. propose regular and immediate feedback as a tool to evaluate risk assessment methods [7]. In such a way, it could be detected that a model is not realistic and may have left out some crucial influence factors.

Phase 2: Rating of the influence factors

After the influence factors are identified, they have to be estimated and rated. This is typically done using an ordinal scale defined by the used method or standard or has some internal company or domain-specific definition. In either way, it comes down to deciding for a class on an ordinal scale which the factor corresponds to in order to be able to rank the factors and use them for later risk comparison. In quantitative methods, this is done differently: here, an actual value, range, or even distribution on a ratio scale, which represents the reality, is chosen (no classification, just estimation of real values including the respective uncertainty). The assignment to classes in qualitative methods is one of the most discussed

problem areas in literature. We will go through the problems and show how quantitative methods can cope with most of them:

(E) Ordinal Scales: Qualitative Risk assessment methods mostly use ordinal scales. According to Stevens [22], ordinal scales only allow for ordering or ranking the items. Therefore arithmetic operations like addition or multiplication are undefined. Nevertheless, in risk assessment methods, this is done nearly all the time without question. Stevens defines the following scales, and their respective defined operations [22]:

- **Nominal scale**: Defines equality or inequality ($=, \neq$) of items. Examples: Different kinds of fruits like oranges, apples, or pears.
- **Ordinal scale**: Defines ordering relations ($<, >$) amongst items. Examples: School grades or the ranks in sports events.
- **Interval Scale**: Defines sums and differences ($+, -$) in addition to the ordering. Examples: Temperature; Time; often, values in sports events are stated as time differences compared to the first place.
- **Ratio scale**: Defines absolute ratios ($*, /$) between items in addition to the difference and ordering relations. Examples: Distance, Weight, Probabilities

A problem here is that by transforming quantitative values into a domain and scale, which only supports ordering relations, we lose the ability to do reasonable arithmetic, estimate uncertainty, or do any sophisticated mathematical analysis. Although the so-called "semi-quantitative" scales may give the illusion of doing calculations, the numbers are just placeholders for the class labels. They do not have mathematical foundations or actual connections to the real world. While one would refrain from multiplying "words" like *high risk* and *moderate impact* together, doing this with arbitrarily assigned numbers suddenly seems plausible. For example, if high risk=3 and moderate impact=3, then the risk is $3 \times 2 = 6$, but what is the meaning of 6?.

(F) Semi-Quantitative Scale-Definition: The problems begin with the definition of a semi-quantitative distribution on the ordinal scale. There are many articles on how to design a numeric ordinal scale for use in a semi-quantitative assessment e.g. [23]–[25]. We give a short review of the different options here. What we want to achieve is a mapping from continuous quantitative data to a discrete ordinal scale. First, we have to decide how many ranks the ordinal scale should have and which ranges of values are assigned to which rank. Furthermore, if the ranks should be used for semi-quantitative arithmetics, the ranks must be assigned to numbers.

Decision 1: Number of Ranks: Does the scale have three levels (e.g., high, medium, low), 4, 5, or even 10 or 100 levels? A high number leads to a seemingly continuous scale, while a lower number is more comfortable to judge due to its coarseness [23]. An even number of levels has no neutral state, and therefore the assessment always points into a direction (either lower or higher). Uneven numbers of levels allow for a neutral position in the middle. In addition to that, Hubbard et al., as well as others, found out that people tend to avoid extreme positions [26], [27]. Therefore, it could sometimes be reasonable to add an even more extreme level to an existing scale to outwit the bias of avoiding the most extreme. Using increasing or decreasing numbers, and even how the scale is presented can affect the outcome [27]. A further aspect of this is the next decision is if every factor should have the same number of levels for simplicity's sake, or if they should have a different number of levels to fit the individual factor better. Scientists, like Rensis Likert, have researched the psychological effects of such scales for nearly a century now (he coined the term "Likert-scale"). However, for conciseness, we leave out further psychological debate about psychometric scales and refer to [28], [29] for further information.

Decision 2: Assignment of Quantitative Ranges to Ranks (Distribution): It is important to decide which ranges of values belong to which rank on the ordinal scale. Is this distribution scaled linear or logarithmic? Table II and Figure 6 show different kinds of distribution numerically and graphically, and here we enlist and describe some of the most common ways to define the distribution of ranks:

- *Linear*: Linear-based scales split a value range into equally distributed ranges and assign labels to them. E.g., low, medium, high. Linear-based ordinal scales relate approximately to ratio scales but still have the problem of assigning arbitrary numbers to the value ranges, which dismisses all arithmetic semantics. Sometimes they are inappropriate because, in reality, processes often behave quadratic or even exponential, and we still want to be able to cover small differences for the lower values. A linear scale would have to be very big and unhandy to cope with such behavior (imagine a scale from 0 to 100 in 0.1-interval steps. It would have 1000 different levels, while a logarithmic scale would only have 4).
- *Logarithmic*: a logarithmic scale considers processes covering large ranges while still being able to classify small ranges also, e.g., yearly, weekly, daily; small amounts of money vs. large amounts; 1, 10, 100, 1000; Injury is also very often scaled logarithmic (e.g., AIS scale).
- *Normally Distributed (Gaussian)*: Scales that are arranged like a bell curve to distinguish between the tiny and huge exceptions, while average cases are all put into the same category. A variant of this is to arrange the values inversely, to distinguish the average cases better, but clumping up the extreme cases.
- *Arbitrary (Fitted)*: Another possibility is a scale that is fitted arbitrarily. This can be a domain-specific definition from experts or a mathematical best fit with respect to some specific metric. With an arbitrary fit, it is possible to set the boundaries between distinct areas based on some criterion other than a mathematical distribution. The problem here is that such a fit could be highly subjective and only valid for a specific situation and point in time. One example is the energy labeling legislation in the EU [30]: While in the past the distinctions from A (best) to G (worst) were sufficient, newer technology-enabled lower

TABLE I
DIFFERENT LABELS AND SEMI-QUANTITATIVE NUMBER ASSIGNMENTS FOR ORDINAL RANKS.

Rating	Probability	Frequency	Increasing	Start at 0	Decreasing	Centered	3 Levels	4 Levels	Spaced out	Exponential
Very Low	Remotely	Never	1	0	5	-2		1	2	1
Low	Unlikely	Seldom	2	1	4	-1	1	2	4	2
Medium	As Likely as not	Sometimes	3	2	3	0	2		6	4
High	Likely	Often	4	3	2	1	3	3	8	8
Very High	Certain	Always	5	4	1	2		4	10	16

TABLE II
VARIANTS OF CLASSIFICATION SCHEMES FOR THE RANGE FROM 0 TO 100.

Level	Linear	Logarithmic	Gaussian	Inv. Gaussian
1	$0 \ldots 20$	$0 \ldots 0.01$	$0 \ldots 10$	$0 \ldots 30$
2	$> 20 \ldots 40$	$> 0.01 \ldots 0.1$	$> 10 \ldots 30$	$> 30 \ldots 45$
3	$> 40 \ldots 60$	$> 0.1 \ldots 1$	$> 30 \ldots 70$	$> 45 \ldots 55$
4	$> 60 \ldots 80$	$> 1 \ldots 10$	$> 70 \ldots 90$	$> 55 \ldots 70$
5	$> 80 \ldots 100$	$> 10 \ldots 100$	$> 90 \ldots 100$	$> 70 \ldots 100$

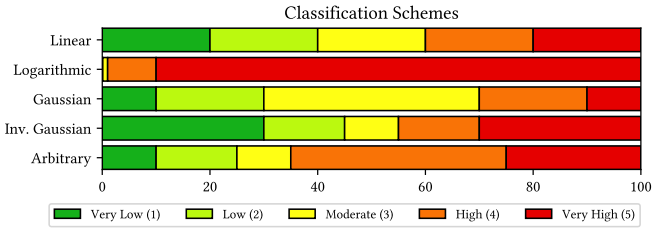

Fig. 6. Illustration of different ranges for classification into ordinal scales.

power consumption, therefore more categories have been introduced over time: A+, A++, A+++, and beginning with 2020 the scale will be completely rearranged to use the labels again A to G [30]. Energy labels based on a quantitative ratio scale would not suffer this problem (e.g., labels with power consumption in Watt and efficiency in percent, or net and gross power labels). Further examples can be found in literature like Ho et al. [24]. They show and compare the arbitrary definitions of probability scales for words of estimative probability [31], defined in several different standards.

Decision 3: Assignment of Semi-Quantitative Numbers to Ranks: Are the semi-quantitative assigned numbers centered around 0, increasing, or decreasing? Is the 0 included or not? Table II shows some examples of different scales, which are also visualized in Figure 6.

Aside from the distribution of values, the direction and location of the centers are significant [32]. Humans are susceptible and biased towards different orders, and labels in scales [27]. Table I shows different variants of number assignments to ordinal scale levels.

- *Increasing*: Numbering the scale in increasing order. This relates to the notion of "higher numbers result in higher risk".
- *Decreasing*: Scaling the levels with decreasing ranges inverts the meaning. Here, less of something corresponds to higher risk, e.g., lower defense means a higher risk of successful attacks.
- *Centered around 0*: Sometimes positive and negative aspects are modeled in the scale. e.g., losses or gains. In such scales, the neutral element is 0, while the more extreme cases fall to either side of the number range (positive or negative).
- *Including or omitting 0*: If the scale includes 0 and the combination includes multiplication, this 0-level could completely wipe out all other properties, regardless of how high they are. This is unwanted behavior since it conflicts with the monotonicity and relevance criteria for coherent risk metrics [8].

All these decisions are somewhat arbitrary and made mostly for convenience to have a more straightforward combination and ranking strategy later on.

***(G) Range Compression**:* Through pressing the real values into a scheme of ordinal scales, the original uncertainty ranges get lost, and the whole value range of an ordinal class is applied to the values. Overlapping ranges get clipped, smaller ranges get widened.

***(H) Ambiguity**:* The scales are often not defined precisely, and therefore can be argued and judged differently based on the experts' opinion.

What is still light injury, what is already severe injury? Where is the border between once per week and once per month? How do *"very low chance"* and *"remote chance"* differ from each other? [24], [33]

***(I) Neglecting Uncertainty**:* By classifying, the original uncertainty in the judgment gets lost. The class imposes a new default range for the uncertainty. This relates strongly to range compression and quantification errors. If the uncertainty was huge and would span multiple classes, this cannot be encoded. If the uncertainty is smaller and would span only a small fraction of a class, this cannot be encoded either and gets lost in the process.

***(J) Quantification Errors**:* Especially on the border, quantification errors can happen. If a value changes slightly, it could step up to the next level in the ordinal scale or fall to the lower level. This could change the result tremendously (imagine going from 2 to a 1 with a multiplied combination, resulting in half of the resulting risk, but in reality, the value just changed a little bit).

***(K) Human Bias**:* Humans are very biased [10], [34], [35]. They are scared of bad outcomes and tend to underestimate the probabilities, or they are biased towards the other way and tend to overestimate bad outcomes (risk affinity bias). Also, humans tend to judge events based on their own

experience, which is, by all means, also very flawed. Also, the cultural background and the daily condition play a huge role (see human inconsistency). This flaw can partly be covered by training for consistency and training for neutrality but is still there. Even if the probabilities are exactly defined, humans tend to misinterpret them [33]. Also, centering bias happens in this phase: Humans like to avoid the extreme values of a scale [26].

(L) Human Inconsistency: It is proven that humans are biased due to anchoring and framing. They change their judgments for the same questions based on the daily condition, the immediate situation before the judgment, or the scaling they have to do.

Phase 3: Combining the ratings

After all the influence factors were rated according to the ordinal scales, they get combined. Most methods do this either multiplicative or additive, some have a weighting scheme for addition, and some also use a deduction factor for reducing the final score. All these methods have no mathematical foundation since ordinal scales do not define arithmetic operations (only ordering relation). Still doing it introduces many problems which are discussed here.

(M) Undefined Semi-Quantitative Arithmetics: As already mentioned, ordinal scales do not support operations like addition, subtraction, multiplication, or division. They only define an ordering relation for ranking them. Any semi-quantitative calculation with such ordinal scale levels is just an arbitrary approach, without any foundations or support from mathematics. Connecting to the example already discussed in problem (E): Ordinal scales - what is the meaning of multiplying two different classes of ratings? Can the words "high risk" and "low severity" be multiplied? No. However, we tend to believe that semi-quantitative numbers can. If a high risk corresponds to 3, and a low severity corresponds to 1, the result would have been $3 \times 1 = 3$. However, here we stepped over a fallacy because if we tried this with the corresponding textual labels, it is clear that this is an invalid operation (high risk $\times$ low severity = ?).

(N) Arbitrary Combinations: The way ratings are combined invalid and undefined, but the actual operations are also chosen arbitrarily. Should we multiply or add up all ratings? Should an optional reduction be subtracted, or should the scale be inverted to have a more comfortable mathematical formula? Should we weigh the ratings before we add them together? How are the weights defined? This strongly depends upon the scale definition (see the problem (F)). On ordinal scales, there is no correct way to do this. It is just a convention or definition by some standard. Mostly the kind of combination is chosen to result in a nice number to judge the final risk. Figure 7 shows all additive combination possibilities of the HARA risk scores in the ISO 26262 [15], compared to the same scores when multiplied. By adding them, there are equal groups of levels used for further processing - but when multiplying, they do not group up that nicely, especially the border between categories QM, A, and B are not intuitively

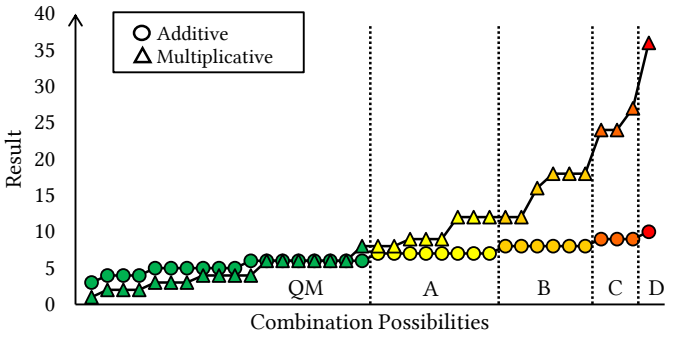

Fig. 7. Comparison of additive and multiplicative combinations of values. The example is based on the Hazard and Risk Analysis, and the areas are the respective ASIL classifications defined in the ISO 26262 [15]. While ASIL-C and ASIL-D are clearly distinguishable in both approaches, the boundaries of QM, A, and B are not that intuitive.

Probability \ + Severity	1	2	3
1	2	3	4
2	3	4	5
3	4	5	6
4	5	6	7

Probability \ * Severity	1	2	3
1	1	2	3
2	2	4	6
3	3	6	9
4	4	8	12

Fig. 8. The tables show the results adding factors compared to multiplying them.

recognizable anymore. In addition to that, the table in Figure 8 shows numerical examples for the combination via addition and multiplication. We can see that multiplying produces more variance and also more risk classes than the addition.

Suppose the ratings include 0 as a number. In that case, multiplication could completely override all other ratings for a risk event, no matter how extreme they are (essentially making a risk irrelevant).

The definition of the weights and combination of influence factors for calculating the risks has developed into an industry because this is needed for actuaries, insurances, finance companies, and clinical and pharmaceutical industries. They try to define their weights and combinations according to some sophisticated model because, for them, it is the basis of million-dollar decisions. One example of a sophisticated combination is the algorithm for calculating the risk of unemployment in Austria's unemployment office [16]. It is a weighted additive scheme based on several demographic and social factors. The weights for combining the factors were inferred via historical data and statistics and reflected society's bias, which was highly disputed in the media. For example, female candidates had a higher risk of staying unemployed than male candidates, or that education was only a minor factor in getting a job.

(O) Dominating Components: If one property on the ordinal scale is low or high, it could dominate the others. For actual ratio scales, this is normal and reasonable. However, since ordinal scales lost their original real-world semantics and the numbers for the scale levels are just arbitrary definitions, the combination is not reasonable anymore. This could be a problem, especially if the scales have different distributions (one is linear, while the other is logarithmic), e.g., due to the

ordinal scales, a level from a linear distribution gets the same influence as a level from a logarithmic distribution. Increasing or decreasing the linear level would result in linear response of the actual risk while increasing/decreasing the logarithmic level would change the risk by a factor of magnitude. However, the calculated ordinal risk metric would still change only by a linear term, regardless of the actual quantitative risk change. This contradicts the positive homogeneity, and the translation invariance property [8].

(P) Neglection of Correlations: One of the most overlooked problems of risk matrices is the neglection of correlations. Cox et al. [3] stated that uncorrected negative correlations between risk matrices' influence factors could lead to worse than random results. While problem (B) already describes this, here, the actual effects are manifested. If we would know the correlations between the input factors, we could correct them in this phase by applying a correlation matrix or some other conversion factor to make up for this. This, of course, would only be possible when we had used quantitative risk assessments with ratio scales. On ordinal scales, it is difficult to model the effect one factor has on another. In the quantitative world, one can detect that a factor changes whenever another factor changes and how they are related to each other (positively or negatively correlated). Since ordinal scale levels are so coarse, this cannot be detected or corrected. The consequences of this may be severe: Correlations could add up and result in a high-risk value, or they annihilate themselves, and the actual risk would not change even when the input factors change. All in all, the calculated risk metric does not reflect the changes in the real quantitative risk, which is a severe problem and contradicts the positive homogeneity property [8].

Phase 4: Ranking and Ordering the risks according to the resulting risk metric

In this phase, the combined risk scores get ranked again and ordered for their importance. As already was the case in phase 2, this ranking is again an ordinal scale and suffers from the same problems. Here, it is even worse because the source data is not a ratio scale but a combined ordinal score which drags along all the problems described until now.

In this phase, the combined risk score is again ranked on an ordinal scale, e.g., all values above a specific threshold get a high rank, all under a specific threshold a low rank, and all in between get medium. E.g., scores from 1 to 5 get a low ranking, scores from 6 to 10 get medium, and 11 to 15 get a high ranking. In addition to all discussed problems of ordinal scales, this phase has even more problems, partly since this final ranking is the basis for decision making.

(Q) Arbitrary Thresholds: The thresholds for the ranges of the final risk levels are often chosen completely arbitrarily, with a high emphasis on simplicity. In the hazard and risk analysis, for example, all scores until 6 are grouped to the lowest risk level (QM), and above 6, every whole integer represents an own risk level (7=ASIL A, 8=ASIL B, 9=ASIL C, 10=ASIL D) [15]. This convention is convenient due to the combination method of addition. This makes it easy to estimate the final risk score already during the individual scores in phase 2.

3x3	0-33	33-66	66-100
0-33	L	L	L
33-66	L	M	M
66-100	L	M	H

4x4	0-25	25-50	50-75	75-100
0-25	L	L	L	L
25-50	L	M	M	M
50-75	L	M	H	H
75-100	L	M	H	H

5x5	0-20	20-40	40-60	60-80	80-100
0-20	L	L	L	L	L
20-40	L	L	L	L	L
40-60	L	L	L	M	M
60-80	L	L	M	M	H
80-100	L	L	M	H	H

Fig. 9. The only possible consistent assignment of risk matrices with linear input scales and a 3-rank output score (L=low, M=medium, H=high), for 3x3, 4x4, and one of the two possible colorings for 5x5, according to Cox [3].

In comparison, quantitative methods also define an arbitrary threshold in this phase, but this would consist of a distribution called the "risk appetite", which defines how much risk in terms of probabilities and real quantitative values is tolerable. For example, how much money loss is still tolerable with a 10% probability, 50% probability, or 90% probability that the loss is realized.

(R) Inconsistency: In their work, Cox et al. [3] describe what consistency for a risk matrix means and why this is important to achieve that property. At the same time, they prove that full consistency cannot be achieved when ordinal scales are used. Consistency means that the resulting risk score should relate directly to the real quantitative risk. For example, it should not be possible to switch from the lowest risk category to the highest by doing just a small change during the evaluation. It should not be possible that actual higher risks get scored below actual lower risks. Furthermore, all events in the same final risk class should represent the same level of actual risk, no matter how they are ranked, combined, and judged during the risk assessment. Cox defines three properties to ensure this: weak consistency, betweenness, and consistent coloring. Figure 9 shows examples for consistent risk matrices using linear scales as input and having 3 ranked ordinal scales as output (high, medium, low).

- Weak Consistency: This property defines that all events which are in the lowest-ranked risk class should have lower actual risks than all the events ranked in the highest risk class. If these two classes are disjunctive, a risk matrix has at least weak consistency.
- Betweenness: It should not be possible to jump directly

from the lowest risk class to the highest by just doing small changes in the input factors. Therefore, at least another class needed to create a border between the lowest and highest classes. This "middle" level may overlap with the lowest or highest class, but it should contain events higher than the lowest class and mostly lower than the highest risk class.
- Consistent Coloring: This property ensures that events on the same risk level should represent approximately the same actual risk in reality. It should not happen that two events are grouped into the same risk class but have opposite actual risks.

(S) Incoherence: Coherence is the notion of general properties for well-behaved risk metrics and was proposed by Artzner et al. [8] in 1999. They argue that a risk metric has to satisfy several properties (or axioms, as they called it) to be useful. Together these properties ensure that a risk metric is reasonable and well behaved. Since the final risk metric is obtained via a risk matrix, it should also satisfy these coherence properties:

- Relevance: $X > 0 \implies p(X) > 0$
  When an event has an actual quantitative risk, the risk metric should also assign some positive value (the risk metric must not be zero). For ordinal scales which exclude the 0, this property holds.
- Monotonicity: $X \geq Y \implies p(X) \geq p(Y)$
  If a real event has a higher risk than another, the risk metric should also come up with a higher or at least the same values. This is already sometimes violated due to the classification into ordinal scales. Using ordinal scales, an event with lower risk might get a higher score than an event with actual higher risk. Just recall the example of oil leakage from the motivation section.
- Translation Invariance: $p(X + \alpha r) = p(X) - \alpha$ This means that, by making some additional effort to reduce the risk, the respective risk metric should decrease by a corresponding amount. It should not happen that increasing or decreasing a risk produces an incoherent change of the risk metric. This also implies that if some action reduces multiple risks by the same amount, their relative order to each other must not change.
- Subaddidivity: $p(X + Y) \leq p(X) + p(Y)$ When combining two risk events, the risk metric should be at most the addition of the single risk metric values. If the events overlap or are somehow correlated, it is less than the sum of the individual values.
- Positive homogeneity: $p(\lambda X) = \lambda p(X)$ This property ensures that the risk metric reflects affine changes in risk. If the risk doubles, also the metric should double.

Real quantitative methods would support these properties already with the most basic risk equation: $Risk = Impact \times Probability$. This equation fulfills all the mentioned properties of coherence and consistency when used with ratio scales or probability distributions.

(T) Ambiguous Order: If the risk matrix is at least weakly consistent, the highest and lowest-ranked risks can be ordered, but what about ordering inside the classes? If different risks result in having the same score, they cannot be prioritized anymore. Furthermore, the middle classes may partially have the same quantitative risk as the lowest or highest classes, making the ordering not very intuitive. An event with a middle-classed score may have a higher actual risk than the higher class score. Also, due to range compression, the highest risks get all clumped up together in the highest class, but the differences could be orders of magnitudes apart.

(U) Risk Inversion: The problem of risk inversion is a very severe one. Lower risks might get a higher score than actually higher risk or vice versa. We will repeat the thought experiment from the motivation section again to make this clear:

Think of an assessment of environmental contamination. Two means of transportation are compared: a car which leaks half a liter of oil every week, and a plane which happens to leak 100 liters every half a year. Furthermore, imagine the following scale for oil leakage:

- $0 \ldots 0.1$ liter = Low impact (1)
- $0.1 \ldots 1$ liter = Medium impact (2)
- $1 \ldots 10$ liters = High impact (3)
- > 10 liters = Very high impact (4)

and the following time scales for the frequency:

(4) Daily = Very high frequency (4)
(3) Weekly = High frequency (3)
(2) Monthly = Medium frequency (2)
(1) Yearly = Low frequency (1)

The car would get a medium impact (2) and high frequency (3), which would result in a risk score of 6, but leaks about 100 liters per year. The plane would get very high impact (4), but only a low frequency (1), which results in a risk score of 4, and the quantitative leakage is 200 liter per year, which is double the leakage of the car. This is a typical example of risk inversion, where an actual higher risk event occurs to be scored with a lower risk score. Even if the plane's frequency rating would have been medium (2), the score would still only be a little higher than the one of the car, while it should be double as high if it really would represent the actual risk.

(V) Unknown Uncertainty and Confidence: We already established that neglecting the uncertainty in phase 1 was not a good idea. Here we have to pay the price for this. Range compression and quantification errors have additionally contributed to completely wipe out any notion of uncertainty or confidence in our data. We cannot determine the uncertainty in our data anymore and are left only with the given ranges coming from the scale levels through the risk scores alone. Maybe our estimations have been very uncertain and may, therefore, be wrong? On the other hand, if we were very confident in our estimations, the scales' ranges and the arbitrary calculations increased the error and uncertainty. How can we ever know if we neglected them along the way?

Quantitative methods could accomplish this by propagating the uncertainty throughout all calculations, or even better, by

using the values' distributions to consider even more details of the underlying quantitative data.

Phase 5: Making decisions based on the Risk Assessment

Based on the scoring and ranking of the risks, we want to decide which ones we want to mitigate and which ones we can tolerate. This last step strongly relates to decision theory [36]. Here, human bias plays a huge role again: the simple traffic-light system of a risk matrix is very appealing for management people. Also, just the task of talking about risks already gives an impression of achievement and benefit. Nevertheless, Hubbard et al. [7] debate that this impression may be deceitful and is just a perceived impression, not a real one. Simply discussing risks may already induce some satisfaction and the notion of accomplishment, but, as Hubbard argues, to make sure that the methods are beneficial, we have to measure their performance. Unfortunately there are little to no evidence that qualitative risk matrices work [9], and even that is just a pure argumentative one, but there are many pieces of evidence that they have quite some problems [3], [4], [6], [7], [32]. One big problem regarding the measurement of performance is that there could be years until some risk eventually occurs - hence there is no immediate feedback, which could be measured easily.

(W) Wrong Impression of Benefits: Because the risk assessment based on semi-quantitative methods seems so "easy" and "natural", there is the notion that it is correct and trustworthy. Risk matrices are established tools, which companies have used for many decades now. They may even appear to be "authoritative, and intellectually rigorous" [32] due to their seemingly correct semi-quantitative approach. However, as we established in this work, this is not the case. The benefit could be just an illusion, again bred by the human bias of uncertainty aversion and authority bias [10].

(X) Deferred Feedback: Hubbard et al. [6], [7], and others [37], stated the actual fact, that immediate feedback is an absolute must for being able to improve. The longer the feedback loop endures, the weaker the learning effect is. For risk assessment methods, the time frame between the assessment and the actual risk event may be years apart. Therefore, the initial evaluation is seldom reviewed for correctness, and methods themselves are even more rarely approved for their validity or performance. Often, the people who did the assessment already left the company long before, making it even more difficult to reevaluate and improve on the estimations. Unfortunately, this is also a problem that applies even to quantitative methods. It is important to check the validity of methods by measuring their prediction strength and comparing this with other methods to find the most suitable method for a purpose.

IV. Conclusion

In this work, we discussed many aspects and problems of risk matrices. We showed that in every phase of the risk assessment process, risk matrices have flaws and may introduce errors that could lead to wrong decisions in the end. By showing this, we made another case against qualitative or semi-quantitative risk assessment methods and proposed quantitative approaches. In our research group, we are currently developing such a method based on quantitative risk assessment for cyber-security, called RISKEE [12]. The mathematics behind it can be used for any risk assessment, and we will make it available as soon as we have enough evidence supporting the correctness. In the future, we plan to investigate problems that exist even when using quantitative methods, e.g., detecting incompleteness or irrelevance of input factors and tackling the problem of deferred feedback to evaluate the appropriateness of the method. Also, combining several different expert judgments to get a realistic judgment is an area we want to tackle in future papers.

Our plea is to the safety and risk experts out there to reflect on the possible pitfalls of risk matrices and review their methods and estimations, whether they may have fallen into some of the possible traps luring inside risk matrices. Furthermore, we encourage using quantitative risk assessment methods wherever possible.

References

[1] A. L. Cox, D. Babayev, and W. Huber, "Some Limitations of Qualitative Risk Rating Systems," *Risk Analysis*, vol. 25, no. 3, pp. 651–662, 2005.

[2] A. L. Cox and D. A. Popken, "Some Limitations of Aggregate Exposure Metrics," *Risk Analysis*, vol. 27, no. 2, pp. 439–445, 2007.

[3] A. L. Cox, "What's Wrong with Risk Matrices?" *Risk Analysis*, vol. 28, no. 2, pp. 497–512, Apr. 2008. [Online]. Available: http://doi.wiley.com/10.1111/j.1539-6924.2008.01030.x

[4] L. A. Cox, *Risk analysis of complex and uncertain systems*, ser. International series in operations research & management science. New York: Springer, 2009, no. 129, oCLC: ocn276223899.

[5] L. A. T. Cox, "Confronting Deep Uncertainties in Risk Analysis," *Risk Analysis*, vol. 32, no. 10, pp. 1607–1629, Oct. 2012. [Online]. Available: http://doi.wiley.com/10.1111/j.1539-6924.2012.01792.x

[6] D. W. Hubbard, *How to measure anything: finding the value of intangibles in business*, 3rd ed. Hoboken, New Jersey: John Wiley & Sons, Inc, 2014, 00000.

[7] D. W. Hubbard and R. Seiersen, *How to measure anything in cybersecurity risk*. Hoboken: Wiley, 2016, 00000.

[8] P. Artzner, F. Delbaen, J.-M. Eber, and D. Heath, "Coherent Measures of Risk," *Mathematical Finance*, vol. 9, no. 3, pp. 203–228, Jul. 1999. [Online]. Available: http://doi.wiley.com/10.1111/1467-9965.00068

[9] Julian Talbot, "What's right with risk matrices? | Julian Talbot on Risk, Success and Leadership," 2018. [Online]. Available: https://www.juliantalbot.com/single-post/2018/07/31/Whats-right-with-risk-matrices

[10] D. Kahneman and A. Tversky, "Subjective probability: A judgement of representativeness," *Cognitive psychology*, vol. 3, no. 3, pp. 430–454, 1972. [Online]. Available: http://datacolada.org/wp-content/uploads/2014/08/Kahneman-Tversky-1972.pdf

[11] J. Dobaj, C. Schmittner, M. Krisper, and G. Macher, "Towards Unified Quantitative Integrated Security and Safety Risk Assessment," *Under review*, p. 14, 2019.

[12] M. Krisper, J. Dobaj, and G. Macher, "RISKEE: A Risk-Tree Based Method for Assessing Risk in Cyber Security," *Under Review*, p. 12, 2019.

[13] USA Department of Defense, "Military Standard MIL-STD-882c - System Safety Program Requirements," 1979.

[14] IEC, "IEC 60812: Analysis techniques for system reliability - Procedure for failure mode and effects analysis (FMEA)," 2006.

[15] ISO, "ISO 26262," 2018.

[16] J. Holl, G. Kernbeiß, and M. Wagner-Pinter, "Das AMS-Arbeitsmarkt-chancen-Modell," p. 16, 2018.

[17] C. Lecher, "How Amazon automatically tracks and fires warehouse workers for 'productivity'," Apr. 2019. [Online]. Available: https://www.theverge.com/2019/4/25/18516004/amazon-warehouse-fulfillment-centers-productivity-firing-terminations

[18] J.-L. Martin and D. Wu, "Pedestrian fatality and impact speed squared: Cloglog modeling from French national data," *Traffic Injury Prevention*, pp. –, Jan. 2017. [Online]. Available: https://hal.archives-ouvertes.fr/hal-01557978

[19] C. Jurewicz, A. Sobhani, J. Woolley, J. Dutschke, and B. Corben, "Exploration of Vehicle Impact Speed – Injury Severity Relationships for Application in Safer Road Design," *Transportation Research Procedia*, vol. 14, pp. 4247–4256, 2016. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S2352146516304021

[20] L. Aarts and I. van Schagen, "Driving speed and the risk of road crashes: A review," *Accident Analysis & Prevention*, vol. 38, no. 2, pp. 215–224, Mar. 2006. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0001457505001247

[21] T. DeMarco, *Controlling Software Projects: Management, Measurement, and Estimates: Management, Measurement and Estimation*. Englewood Cliffs, N.J: Pearson Education, Nov. 1982.

[22] S. S. Stevens, "On the Theory of Scales of Measurement," *Science*, vol. 103, no. 2684, pp. 677–680, 1946.

[23] T. R. Knapp, "Treating Ordinal Scales as Interval Scales: An Attempt To Resolve the Controversy," *Nursing research*, vol. 39, no. 2, pp. 121–123, 1990.

[24] E. H. Ho, D. V. Budescu, M. K. Dhami, and D. R. Mandel, "Improving the communication of uncertainty in climate science and intelligence analysis," *Behavioral Science & Policy*, vol. 1, no. 2, pp. 43–55, 2015.

[25] E. D. Smith, W. T. Siefert, and D. Drain, "Risk matrix input data biases," *Systems Engineering*, vol. 12, no. 4, pp. 344–360, Sep. 2009. [Online]. Available: http://doi.wiley.com/10.1002/sys.20126

[26] D. W. Hubbard, *The failure of risk management: why it's broken and how to fix it*. Hoboken, N.J: Wiley, 2009, oCLC: ocn268790760.

[27] G. Moors, N. D. Kieruj, and J. K. Vermunt, "The Effect of Labeling and Numbering of Response Scales on the Likelihood of Response Bias," *Sociological Methodology*, vol. 44, no. 1, pp. 369–399, Aug. 2014. [Online]. Available: http://journals.sagepub.com/doi/10.1177/0081175013516114

[28] N. K. Malhotra, *Basic Marketing Research*, 4th ed. Boston: Pearson, Jul. 2011.

[29] J. Dawes, "Do Data Characteristics Change According to the Number of Scale Points Used? An Experiment Using 5-Point, 7-Point and 10-Point Scales," *International Journal of Market Research*, vol. 50, no. 1, pp. 61–104, Jan. 2008. [Online]. Available: https://doi.org/10.1177/147078530805000106

[30] European Commission, "REGULATION (EU) 2017/ 1369 OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL - of 4 July 2017 - setting a framework for energy labelling and repealing Directive 2010/ 30/ EU," p. 23, 2017.

[31] S. Kent, "Words of Estimative Probability," 2012. [Online]. Available: https://www.cia.gov/library/center-for-the-study-of-intelligence/csi-publications/books-and-monographs/sherman-kent-and-the-board-of-national-estimates-collected-essays/6words.html

[32] P. Thomas, R. B. Bratvold, and E. Bickel, "The Risk of Using Risk Matrices," in *SPE Annual Technical Conference and Exhibition*. New Orleans, Louisiana, USA: Society of Petroleum Engineers, 2013. [Online]. Available: http://www.onepetro.org/doi/10.2118/166269-MS

[33] P. D. Windschitl and E. U. Weber, "The Interpretation of "Likely" Depends on the Context, but "70%" Is 70%—Right? The Influence of Associative Processes on Perceived Certainty," *Journal of Experimental Psychology: Learning, Memory, and Cognition*, vol. 20, no. 6, p. 20, 1999.

[34] D. D. Ariely, *Predictably Irrational: The Hidden Forces That Shape Our Decisions*. New York, NY: Harper Perennial, Apr. 2010.

[35] T. Gilovich, D. Griffin, and D. Kahneman, *Heuristics and Biases: The Psychology of Intuitive Judgment*. Cambridge, U.K. ; New York: Cambridge University Press, Jul. 2002.

[36] R. A. Howard and A. E. Abbas, *Foundations of Decision Analysis*, 01st ed. Boston: Pearson, Jan. 2015.

[37] C. Newport, *Deep work: rules for focused success in a distracted world*. London: Piatkus, 2016, oCLC: ocn951114416.