Aalto University
School of Science
Degree Programme in Engineering Physics and Mathematics

Eliciting Expert Judgements for Probabilistic Cross-Impact Assessment

Bachelor's Thesis
August 27, 2020

Andrea Lyly

Aalto University
School of Science

Aalto University, P.O. BOX 11000, 00076 AALTO
www.aalto.fi
Abstract of the bachelor's thesis

Author Andrea Lyly

Title Eliciting Expert Judgements for Probabilistic Cross-Impact Assessment

Degree programme Engineering Physics and Mathematics

Major Mathematics and System Sciences **Code of major** SCI3029

Teacher in charge Prof. Ahti Salo

Advisor M.Sc Tech. Juho Roponen

Date 27.8.2020 **Number of pages** 31+13 **Language** English

Abstract

Expert judgements are used to understand uncertain events of the present and future when data is not available. The purpose of this bachelor's thesis was to develop a process for eliciting expert judgements that can be used for risk assessment purposes. The efficiency and accuracy of the process was tested in an experimental case study conducted with a small panel of five experts.

We are interested in evaluating expected risks of systems using scenarios. Scenarios are combinations of outcomes that come from uncertainty factors. Complex systems usually have many uncertainty factors with various outcomes, and this can lead to a massive number of scenarios. Therefore, trying to estimate scenario specific probabilities one by one is just not practical. Instead, in order to reduce workload, we use probabilistic cross-impact analysis, which requires eliciting expert judgements on the marginal probabilities and pairwise relationships of the outcomes. Obtaining and analyzing data is done by using methods from the Cooke's Classical Model, which aims in evaluating the judgements of the experts according to their calibration scores and informativeness. The judgement elicitation process is based on the Delphi method.

The results in the case study showed that the process worked efficiently and in a correct way, although, the process was performed with remote connections. Giving judgements for the marginal probabilities was considered to be easy among the experts, and the judgements yielded in reliable probabilistic data. Judging the relationships of the pairwise outcomes was considered to be also fairly effortless, but the mathematical analysis faced some challenges as the judgements were initially non-feasible. However, the method used in the thesis helped to obtain a feasible solution by slightly correcting the judgements of the experts. The calibration results for the pairwise relationship judgements were not the best, and therefore the judgements were not considered to be reliable, which also affects the reliability of the risk assessment.

We also discuss how the process could be improved. In larger scale studies, the process is likely to take much more time and effort, and performing the process remotely becomes challenging.

Keywords eliciting expert judgements, probabilistic cross-impact analysis, risk analysis

Aalto University
School of Science

Aalto-yliopisto, PL 11000, 00076 AALTO
www.aalto.fi
Tekniikan kandidaatintyön tiivistelmä

Tekijä Andrea Lyly

Työn nimi Asiantuntija-arvioiden määrittäminen todennäköisyyspohjaiseen ristivaikutusarviointiin

Koulutusohjelma Teknillinen fysiikka ja matematiikka

Pääaine Matematiikka ja systeemitieteet **Pääaineen koodi** SCI3029

Vastuuopettaja Prof. Ahti Salo

Työn ohjaaja DI Juho Roponen

Päivämäärä 27.8.2020 **Sivumäärä** 31+13 **Kieli** Englanti

Tiivistelmä

Asiantuntija-arvioita käytetään nykyhetken ja tulevaisuuden epävarmojen tapahtumien ymmärtämisessä, kun muuta tietoa ei ole saatavilla. Tämän kandidaatin tutkielman tarkoituksena oli kehittää prosessi, jolla määritetään asiantuntija-arvioita riskiarviointeja varten. Prosessin tehokkuutta ja tarkkuutta testattiin kokeellisessa tutkimuksessa, joka toteutettiin viiden asiantuntijan kanssa.

Riskejä voidaan arvioida skenaarioiden avulla. Skenaariot määritetään epävarmuustekijöiden toteumien yhdistelminä, jolloin monimutkaisissa systeemeissä skenaarioiden määrä saattaa olla valtava. Tällöin skenaariokohtaisten todennäköisyyksien arvioiminen yksitellen ei ole käytännöllistä. Sen sijaan työmäärän vähentämiseksi käytämme todennäköisyyspohjaista ristivaikutusanalyysiä, johon määritämme asiantuntijoiden arvioiden avulla epävarmuustekijöiden toteumien reunatodennäköisyydet sekä toteumaparien suhteet toisiinsa nähden. Arvioiden määrittäminen ja analysointi suoritetaan käyttämällä menetelmiä Cooken klassisesta mallista, jonka tavoitteena on painottaa asiantuntijoiden arvioita heidän kalibrointitulosten ja informatiivisuuden perusteella. Prosessin toiminta perustuu Delfoi-menetelmään.

Kokeellisen tutkimuksen tulokset osoittivat, että prosessi toimi tehokkaasti ja oikein, vaikka tutkimus suoritettiin etäyhteyksillä. Reunatodennäköisyyksien arvioimisen katsottiin olevan helppoa asiantuntijoiden keskuudessa, ja ne tuottivat luotettavaa dataa. Asiantuntijoille toteumaparien suhteiden arviointi oli myös melko vaivatonta, mutta niiden matemaattisessa analyysissä kohdattiin haasteita, koska arviot eivät olleet aluksi matemaattisesti loogisia. Työssä käytetty menetelmä kuitenkin auttoi toteuttamiskelpoisen ratkaisun löytämisessä korjaamalla hieman asiantuntijoiden arvioita. Kalibrointitulokset toteumaparien suhteiden arvioinnissa eivät kuitenkaan olleet parhaita, joten arvioita ei pidetty kovin luotettavina, mikä vaikuttaa myös riskinarvioinnin luotettavuuteen.

Työssä pohditaan myös miten prosessia voisi parantaa. Laajemmissa tutkimuksissa prosessiin kuluu todennäköisesti paljon enemmän aikaa sekä vaivaa, ja prosessin suorittaminen etäyhteyksillä muuttuu haastavaksi.

Avainsanat asiantuntija-arviot, todennäköisyyspohjainen ristivaikutusanalyysi, riskiarviointi

Contents

1 Introduction

Predicting scenario based risks of the present moment and future is an important part of operational research in making better decisions, and these are applied in various fields, such as military planning, nuclear waste reposition and natural hazard management. However, predicting scenarios is very difficult, because the future events in the society, nature, and companies are often uncertain (Seeve, 2018). Furthermore, data is not always available to forecast events, and therefore judgements provided by experts to quantify uncertainty of the future are crucial. Expert opinions and judgements help in decision making and clarifying results (Keeney and von Winterfeldt, 1989), and acquiring these judgements is what we call elicitations.

Eliciting expert judgements is not trivial. Elicitation requires subject knowledge, both from the expert and the analyst conducting the elicitation. Judgements themselves are not sufficient enough for forecasting the future, and therefore we use the Cooke's Classical Model presented in Dias et al. (2018) in order to mathematically interpret and validate the judgements. In addition, Oakley (2010) brought up that psychological aspects play an important role in elicitation. Furthermore, even if the expert judgements are mathematically valid, they still need to be interpreted critically, because there are always possibilities that the judgements are biased on some level (Dias et al., 2018; Kynn, 2008).

The objective of this thesis is to develop an efficient and accurate process for eliciting expert judgements. The judgements are intended to be used for risk assessment purposes using the probabilistic cross-impact analysis (PCIA) presented in Salo et al. (2020). Specifically, this thesis seeks to combine the methodologies from previous elicitation models in Dias et al. (2018) and the new PCIA approach, and assemble them into a single process. The elicitation process is tested in an experimental case study with a small panel of experts.

This thesis contains the following sections. Section 2 explores previous studies related to the subject of the thesis. Section 3 is an introduction to scenario analysis and PCIA. In section 4 we introduce the developed elicitation process and the methodology behind it. The problem formulation and the results for the experimental case study are presented in section 5, where we also analyze the elicitation process. Finally, section 6 concludes the thesis.

2 Background

Expert information is defined in Martin et al. (2012) as information of a topic that is not widely known by others. Furthermore, the experts can give judge-

ments based on their information, which can be simple yes/no answers, thorough descriptions for events or probabilistic data. Expert judgements have been used widely in decision making and operations research, for example, in forecasting weather (Doswell III, 2004), estimating risks of earthquakes (Dias et al., 2018), assigning agents on tasks in order to optimize operative results (Kangaspunta and Salo, 2014) and to obtain qualitative knowledge on relationships of events (Jeong and Kim, 1997). Recent studies have been focused on risk management for safety-critical systems, for example, estimating effects of nuclear waste repositories (Tosoni et al., 2019). Judgements acquired from experts are still used even when data of various phenomena is available, because experts can give opinions that support the mathematical aspects of an analysis and give the results a human-factor interpretation.

A process to acquire valid expert judgements is called an elicitation. Dalkey and Helmer (1963) introduced and developed one of the most commonly used elicitation method, the Delphi method, which aims in quantifying group opinions of a panel of experts by eliciting judgements individually from experts through questionnaires and enhance the experts ability to forecast future events by giving controlled feedback from each round of the process. Avella (2016) summarised that the main advantages of the Delphi method lie in minimizing group bias, i.e. every expert can express their own judgements without being suppressed by others. He also states, the method is flexible, cost efficient and can be performed remotely. However, disadvantages for the Delphi method are also present, as he summed up that the main flaws are caused by the researchers conducting the elicitation. For example, how the questionnaires are made and the conditions on how the experts are chosen can cause bias in the whole process. Helmer (1977) brought up that the Delphi method has also been criticised for the way it fails to observe properly randomized polling procedures, this is, however, a controversial criticism, because the Delphi method is not an opinion poll, which relies on random samples of the population.

In addition to the Delphi method, many other methods for eliciting probabilistic and quantitative judgements have been developed, and they can be combined with the Delphi method. One of the most common models for eliciting probabilistic data is the Cooke's Classical Model, presented in Dias et al. (2018), which focuses on quantifying uncertainty judgements from experts on already occurred events and mirror the accuracy of these judgements into predicting future uncertainty. Other models for eliciting probabilistic data are, for example, the IDEA (Investigate Discuss Estimate Aggregate) protocol in Hanea et al. (2017) and the SHELF method developed by Oakley and O'Hagan (2019).

Cross-Impact Analysis (CIA) is a method that is designed to compare pairwise relationships of multiple events and determine how they would impact the

resulting events in the future. Thus, the method is used widely in scenario analysis in order to forecast future (Bañuls and Turoff, 2011). Furthermore, CIA can be performed in qualitative approaches, e.g. Jeong and Kim (1997), and also quantitative approches, where the aim is to quantify scenario impacts on political, social, technological and environmental events. For example, Weimer-Jehle (2006) presented a theoretical cross-impact balance (CIB) approach, which aims in identifying most consistent scenarios using pairwise comparisons of consistencies. However, even when the responses are systematically recorded using measurement scales, the cross-impact balance method fails to identify consistent scenarios (Salo et al., 2020). A probabilistic approach to CIA, introduced by Salo et al. (2020), quantifies scenario specific probabilities using probability and consistency judgements on events, in a more strict manner. This allows to evaluate expected risks of safety critical systems and identify most probable scenarios. This methodology is discussed more in section 3.

3 Methodology

3.1 Scenarios

Scenarios are combinations of different uncertain future events called uncertainty factors, which are the key components that drive the change in the operational environment (Seeve, 2018). Each uncertainty factor has outcomes, which can be modeled as discrete or continuous, and they must be defined accurately in order to construct scenarios. This thesis uses the definitions of uncertainty factors, outcomes and scenarios presented in Seeve (2018).

Let Y_i be an uncertainty factor, where $i = 1, 2, ..., n$. An uncertainty factor is defined as a set of outcomes, where the total number of outcomes is denoted by K_i. Let A_i be an arbitrary outcome of Y_i,

$$A_i \in Y_i = \{1, 2, ..., K_i\}.$$

From this information it is possible to define a scenario s, which is a vector containing n number of values, where each value is an outcome from every uncertainty factor:

$$s = \begin{bmatrix} A_1 & A_2 & \cdots & A_n \end{bmatrix} \in S,$$

where S is the set of all possible scenarios, and it is defined as the cross product of all uncertainty factors:

$$S = Y_1 \times Y_2 \times \cdots \times Y_n.$$

The total number of the scenarios N, is the product of the number of outcomes of each uncertainty factor

$$N = \prod_{i=1}^{n} K_i.$$

Thus, the number of the scenarios increases exponentially, when the number of uncertainty factors and their outcomes are increased. The number of scenarios N is crucial in the context of analysing scenarios for risk assessment purposes, because complex systems are formed from multiple uncertainty factors with various outcomes. For example by having 9 uncertainty factors with 5 outcomes each, the number of scenarios is $N = 5^9 \approx 1950000$. Clearly, in this case, estimating probabilities for each scenario one by one is impossible in terms of resources and time if data is not available. Thus, cross-impact analysis is used in order to solve the challenge of having many scenarios.

3.2 Probabilistic cross-impact analysis

Here, we cover the theory of the probabilistic cross-impact analysis (PCIA). In PCIA the main target is to map a joint probability distribution across all possible scenarios (Salo et al., 2020), which can be used in evaluating the expected risk of a system. The idea in PCIA is to obtain information on the marginal probabilities of the outcomes and cross-impact judgements that quantify the relationships of the outcomes. These judgements are interpreted as cross-impact multipliers, which are denoted by C_{ij}^{kl}, where i, j are the indices for the uncertainty factors and k, l are the indices for their outcomes respectfully. The cross-impact multipliers are defined as follows:

$$C_{ij}^{kl} = \frac{P(A_i^k|A_j^l)}{P(A_i^k)}, \tag{1}$$

where $P(A_i^k|A_j^l)$ is the probability of outcome A_i^k given that outcome A_j^l occurs, and $P(A_i^k)$ is the marginal probability of outcome A_i^k. Basically, C_{ij}^{kl} describes how much more likely we are going to observe outcome A_i^k if we assume that A_j^l occurs with certainty.

The cross-impact multipliers are non-negative real numbers $C_{ij}^{kl} \in \mathbb{R}_{\geq 0} = \{x \in \mathbb{R} | x \geq 0\}$, and are defined if and only if both marginal probabilities $P(A_i^k),\ P(A_j^l) \in (0, 1]$. The cross-impact multipliers can be assigned with the value zero if and only if the joint probability of outcomes A_i^k and A_j^l is zero, i.e $C_{ij}^{kl} = 0 \iff P(A_i^k \cap A_j^l) = 0$. In this situation the interpretation of the value $C_{ij}^{kl} = 0$ is that the outcomes A_i^k and A_j^l are mutually exclusive, and therefore cannot occur at the same time. Assigning $C_{ij}^{kl} = 1$, indicates that the two outcomes A_i^k and A_j^l are independent. This is due to the property that the conditional probability $P(A_i^k|A_j^l) = P(A_i^k)$ if and only A_i^k and A_j^l

are independent, resulting in $C_{ij}^{kl} = P(A_i^k|A_j^l)/P(A_i^k) = P(A_i^k)/P(A_i^k) = 1$. Assigning cross-impact multipliers values in the range $0 < C_{ij}^{kl} < 1$, indicates that the occurrence of the outcome A_j^l reduces the probability of outcome A_i^k to occur. On the other hand, when cross-impact multipliers have values $C_{ij}^{kl} > 1$, the occurrence of outcome A_j^l increases the probability of outcome A_i^k to occur.

Only one cross-impact multiplier is required for each pair of outcomes, because cross-impact multipliers are symmetric. This can be proven using the Bayes' theorem:

$$P(A_i^k \cap A_j^l) = P(A_i^k|A_j^l)P(A_j^l)$$
$$C_{ij}^{kl} = \frac{P(A_i^k \cap A_j^l)}{P(A_i^k)P(A_j^l)} = C_{ji}^{lk}.$$

The cross-impact multipliers are stored in a symmetric cross-impact matrix (CIM). Another property of the CIM is that cross-impacts of an event with respect to the event itself are not defined, and therefore can be left out. In Figure 1 is shown a cross-impact matrix, where the black cells indicate cross-impacts with the same uncertainty factor and the grey cells are not needed to be calculated, because they are symmetric with respect to the blue cells.

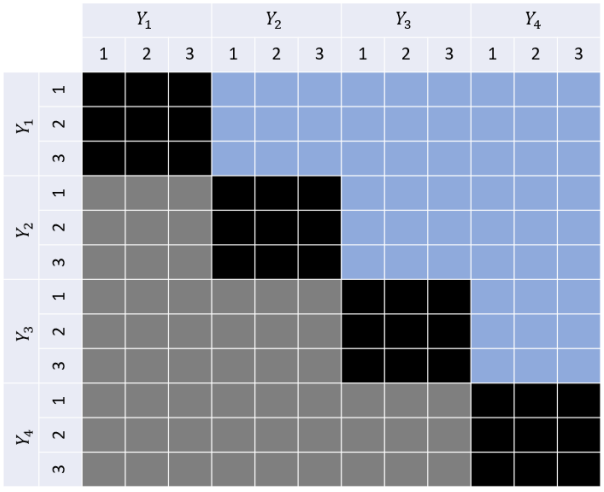

Figure 1: The cross-impact matrix of four uncertainty factors with three outcomes each.

The total number of outcomes N_A, which also translates into the number of marginal probabilities to be elicited, is defined as the sum of the number of outcomes associated with each uncertainty factor: $N_A = \sum_{i=1}^{n} K_i$. By knowing the CIM is symmetric and the diagonal cross-impacts are not defined, we can calculate the number of important cross-impacts N_I of the CIM (blue cells in Figure 1) $N_I = \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} K_i K_j$.

Clearly, is simplistic cases, where the number of uncertainty factors and outcomes are small, we need to estimate more values compared to just estimating the scenario probabilities one by one. For example, having three uncertainty factors with two outcomes each, we end up with $N = 8$ scenarios, but we need to estimate $N_A = 6$ marginal probabilities and $N_I = 12$ cross-impacts. However, as mentioned before, by having 9 uncertainty factors with 5 outcomes each, we end up with $N \approx 1950000$ scenarios. Using cross-impact analysis, we need to estimate only $N_A = 45$ marginal probabilities and $N_I = 900$ cross-impacts, which is significantly more efficient. One thing to keep in mind is that the marginal probabilities and cross-impact multipliers do not map the joint distribution of the system accurately. Obtaining precise probabilities for the scenarios one by one will yield to an accurate joint distribution.

3.3 Risk analysis

In the context of probabilistic risk analysis, we are interested in evaluating expected risk of a system containing multiple scenarios. These scenarios can induce various consequences, which can be interpreted as financial losses, casualties or disutilities. The expected risk is measured in terms of exceeding a specific regulatory threshold level $\zeta \in \mathbb{R}$. Salo et al. (2020) covered the methodology of this following concept. The idea is to denote consequences as a random variable Z, which indicates the consequence of the system. Thus, we can condition the consequence Z on scenarios and obtain a probability of exceeding the regulatory threshold

$$P(Z > \zeta) = \sum_{s \in S} P(Z > \zeta | s) P(s), \tag{2}$$

where $P(Z > \zeta | s)$ is the probability of exceeding the regulatory threshold in the scenario s and $P(s)$ is the probability of the scenario s.

Furthermore, we can generalize the expression (2) by denoting Z_r as unacceptable consequences that belong to a set of unacceptable consequences $\mathbb{C}^{\text{fail}}$. These consequences are not real valued. The following disutility function

$$U(Z_r) = \begin{cases} 1, & Z \in \mathbb{C}^{\text{fail}} \\ 0, & Z \notin \mathbb{C}^{\text{fail}} \end{cases}$$

determines if the consequences belongs to the set $\mathbb{C}^{\text{fail}}$. Using this disutility function we can calculate the probability with which the consequences will be unacceptable

$$E[U(Z_r)] = \sum_{s \in S} E[U(Z_r) | s] P(s). \tag{3}$$

Calculating real valued expected consequences $E[Z]$ is done by estimating the scenario specific expected consequences $E[Z|s]$. Additionally, expected utilities/disutilities $E[U(Z)]$ are calculated by estimating the scenario specific expected utilities/disutilities $E[U(Z)|s]$ with a proper von Neumann-Morgenstern utility function (von Neumann and Morgenstern, 1953). We can apply the information in the following equations

$$E[Z] = \sum_{s \in S} E[Z|s]P(s) \tag{4}$$

$$E[U(Z)] = \sum_{s \in S} E[U(Z)|s]P(s). \tag{5}$$

In the probabilistic risk assessment, we can use one of the equations (2), (4) or (5) as an objective function, which is minimized or maximized, subject to the available information on marginal probabilities and cross-impact multipliers. The proper objective function is selected depending on the target of the risk analysis, and what we are interested in knowing.

By minimizing the selected objective function, we obtain a lower limit of the expected risk, and by maximizing, we obtain the upper limit. According to Salo et al. (2020)

- If the lower limit exceeds the tolerable risk level, the system is deemed unsafe.
- If the upper limit is below the tolerable risk level, the system is deemed safe.
- Otherwise, the safety of the system is uncertain. Revisiting the consequences, estimated probabilities and cross-impacts is justified.

In this thesis, we use a ready-made optimization algorithm used in Salo et al. (2020).

4 Expert elicitation

The elicitation process here follows the Delphi method in Dalkey and Helmer (1963) and Cooke's Classical Model from Dias et al. (2018). Moreover, it contains marginal probability estimation with two different methods, confidence interval calculations, cross-impact multiplier elicitation and feasibility testing. The target of the process is to obtain judgements for the decision maker (DM). The facilitator is in charge of eliciting the judgements from the experts and the analyst uses the elicited data to construct the models for the DM.

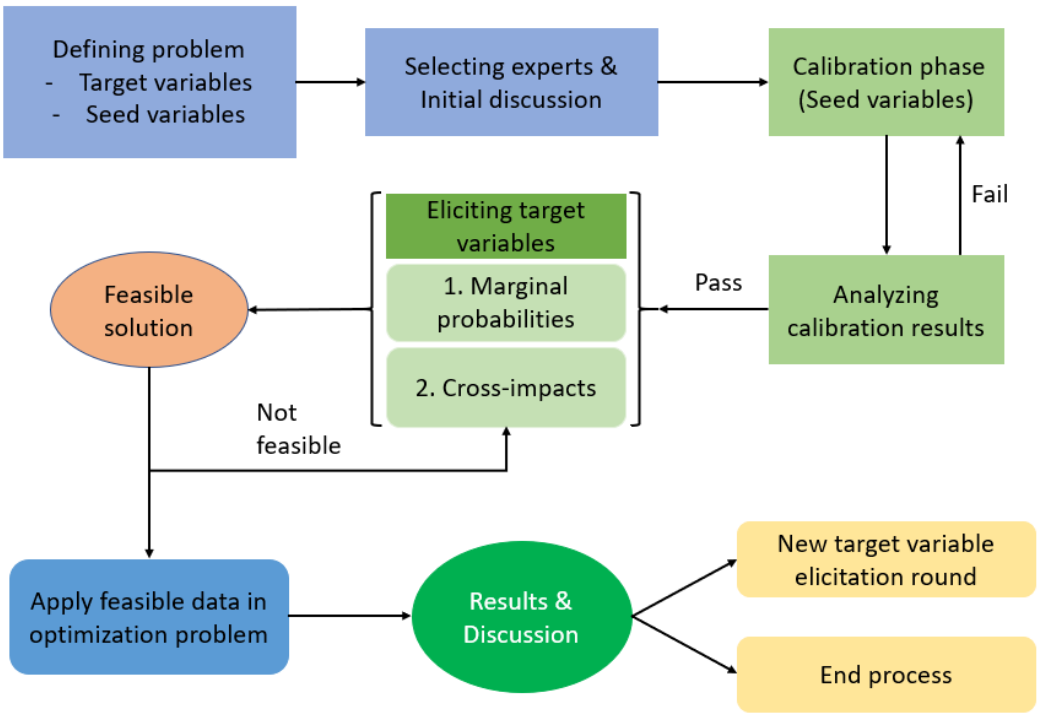

Figure 2: The diagram of the elicitation process.

4.1 Elicitation process

The elicitation process is illustrated in Figure 2, and it begins by defining the problem and identifying uncertainty factors and their possible outcomes. The basic concept in the Classical Model is that it considers two types of questions, regarding target variables and seed variables. The target variables are the variables of interest in the study, i.e. the marginal probabilities of the outcomes and cross-impact multipliers. The seed variables are carefully chosen variables that are closely related to the target variables, where the analyst knows the true value, but the experts do not. The experts are not expected to know the precise true values, however, they are expected to give accurate and informative judgements. Usually, the number of seed variables is 8-20 (Dias et al., 2018). It is important that the questions regarding the seed variables are not something the expert can easily recall from memory, therefore, seed variables can be related to something that has occurred presently or will occur within a short period of time. Moreover, the research on the seed variables and question formulation must be done correctly, because Avella (2016) points out that this phase can cause bias in the process and in the worst case it may affect the overall results.

The next phase is selecting experts. They must be selected by taking into

account their background and reputation on the specific field. An expert does not have to be an expert by job title, but can be a person who is knowledgeable about a specific field. The experts are called together for a starting discussion, where they are briefed on the layout of the problem and the methodology. Furthermore, a lecture in probability theory and critical thinking is provided if needed (Kynn, 2008). Before the calibration phase, the experts are separated from each other to minimize group bias and to allow them to give their own judgements.

Expert calibration quantifies the uncertainty judgements, and it measures the accuracy of their judgements about the seed variables (Dias et al., 2018). If an expert is not well calibrated, then their judgements weigh less on the overall analysis, because the Cooke's Classical Model assumes that the experts perform equally well on the target variables as they performed on the seed variables. Depending on the type of target variables, calibrations are performed with different methods. However, if the number of target variables of one type is small compared to the rest, for example, 10 continuous target variables and 1 discrete, it may not be prudent to perform a calibration for just one discrete target variable. The DM will make a decision on how to proceed with the calibration. In addition, the experts are not specifically told that they are answering calibration questions, because the experts should be unbiased and give judgements purely on knowledge without the pressure of giving good calibrations.

The next phase is the elicitation process is eliciting judgements for the target variables. The methodology's depending on the target variable type are explained in sections 4.2, 4.3 and 4.4. The results of this phase are tested for feasibility in the optimization problem, which is explained in 4.4. If the results are feasible, they are used to analyze the risk of the studied system.

The elicitation concludes with a final discussion with the expert panel. At this point the experts assemble to analyze the obtained judgements and results. If the experts and the DM are satisfied with the results, they can proceed to make a decision. Otherwise, a new round for eliciting judgements for target variables is justified. In the new round, the experts give refining judgements with controlled feedback according to the Delphi method (Dalkey and Helmer, 1963).

4.2 Continuous probability distributions

Calibration for continuous variable elicitation in Dias et al. (2018) is done by collecting quantile estimations on seed variables. Usually, the experts assess the 5th, 50th and 95th quantiles of the seed variables, i.e. the lower

bound, median and upper bound. These quantiles form $t = 4$ intervals shown in Table 1, and the expected proportion of realizations of the intervals are $p = [p_1, p_2, p_3, p_4] = [0.05, 0.45, 0.45, 0.05]$, i.e. 90% of the right answers are expected to be located between the lower and upper bounds. The experts observed proportion of realizations $r = [r_1, r_2, r_3, r_4]$ can be determined by observing how the real answers are placed in the estimations.

Table 1: Observed and expected proportions of calibration questions.

Quantiles	Below 5th	5th to 50th	50th to 95th	Over 95th
Observed	r_1	r_2	r_3	r_4
Expected	p_1	p_2	p_3	p_4

The divergence between the observed and expected proportions is calulated with the Kullback-Leibler divergence measure

$$I_{e_i}(r, p) = \sum_{j=1}^{t} r_j \ln(\frac{r_j}{p_j}). \tag{6}$$

The Kullback-Leibler divergence measure will equal to zero if the observed proportion $r = p$, and therefore smaller divergence measure is better. If the observed proportion has $r_j = 0$ for some j, then the divergence measure at that point is zero, because $\lim_{r_j \to 0^+} r_j \ln(r_j/p_j) = 0$. According to Dias et al. (2018), analysing expert assessments of observed proportions will equal the expected proportions in the long run. Furthermore, the probability distribution of the divergence measure is related to the Chi-squared distribution χ^2 for large sample sizes. More formally

$$Pr\{2qI(r, p) \leq X\} \to \chi^2_{t-1}, \;\; as \;\; q \to \infty,$$

where q is the number of seed variables and χ^2_{t-1} is the cumulative distribution function (CDF) of the χ^2-distribution with $t - 1$ degrees of freedom. The calibration score $\text{Cal}(e_i)$ of the expert e_i, where $i = 1, 2, ..., m$ is calculated from the CDF of the Chi-squared distribution with $t - 1 = 4 - 1 = 3$ degrees of freedom. The calibration score for the expert is defined as the probability of exceeding the divergence measure $2qI_{e_i}(r, p)$:

$$\text{Cal}(e_i) = 1 - \chi^2_3\{X < 2qI_{e_i}(r, p)\} \in [0, 1]. \tag{7}$$

The calibration score ranks the experts in a quantified way. With the calibration score one, the expert is perfectly calibrated and is assumed to estimate the target variables almost perfectly. However, calibration scores can also be very low, and therefore some experts may not be weighted when calculating final probability distributions of the decision maker. A threshold level α is set up at

the beginning of the elicitation, and it indicates what is the lowest calibration score that can be accepted. Typically, α is set around the value 0.01, but it can be changed during the elicitation. An indicator function determines if the expert e_i will not be weighted:

$$1_\alpha(\mathrm{Cal}(e_i)) = \begin{cases} 1, & \text{if } \mathrm{Cal}(e_i) \geq \alpha \\ 0, & \text{if } \mathrm{Cal}(e_i) < \alpha \end{cases} \tag{8}$$

The value one indicates that the expert has passed the calibration and the value zero indicates the expert has not passed the calibration. The set of qualified experts is

$$\tilde{E} = \{e_i \in E \mid 1_\alpha(\mathrm{Cal}(e_i)) = 1\} \subset E.$$

If $\tilde{E} = \emptyset$, the threshold level α needs to be modified or the calibration must be performed again. To clarify, the experts not passing the calibration phase are not removed from the elicitation or considered "bad", they will still answer also the target variable questions in order to bring reference to the other estimations.

In addition to the calibration, we want our experts to also be informative in their estimations. To analyse the informativeness, we calculate an information score for every seed and target variable of the expert e_i. The information score is calculated using the Kullback-Leibler divergence measure as in (6), and the idea is to compare how much the estimations of the expert diverge from the intrinsic range of the uniform distribution (Dias et al., 2018), illustration in Figure 3. Good estimations are the ones that diverge the most from the uniform distribution.

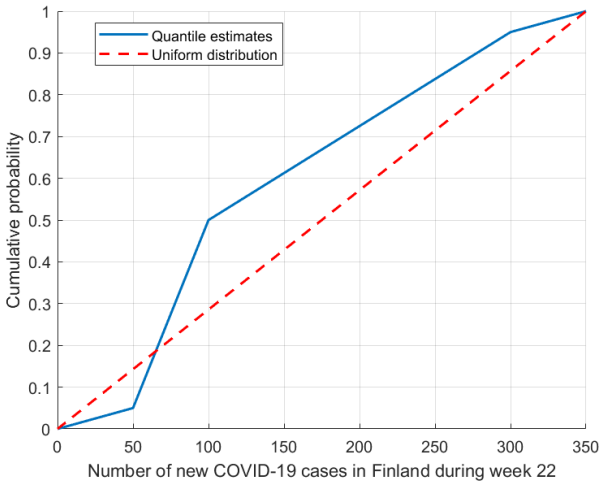

Figure 3: Example of the 5th, 50th and 95th quantiles elicited from an expert compared to the uniform distribution.

Let x_{e_i1}, x_{e_i2} and x_{e_i3} be the estimated 5th, 50th and 95th quantiles of the expert e_i respectfully, and let x_{e_i0} be the minimum value of the seed/target variable and x_{e_i4} the maximum value for the seed/target variable. The minimum and maximum values are the same for all experts e_i, and they can be fixed by the analyst, but usually, the range is determined by the experts. This is done by selecting the expert e_i, who has the biggest difference $d = |x_{e_i3} - x_{e_i1}|$ in his/her 5th and 95th quantiles. The minimum and maximum values are defined as $x_{e_i0} = \min[x_{e_i1}] - \gamma d$ and $x_{e_i4} = \max[x_{e_i3}] + \gamma d$, where γ is an arbitrary proportion (usually 0.1) and $\min[x_{e_i1}]$ is the minimum lower bound of all experts and $\max[x_{e_i3}]$ is the maximum upper bound of all experts. The difference between the maximum and minimum values is defined as $\Delta d = x_{e_i4} - x_{e_i0}$, and this is called the intrinsic range, where the uniform distribution lies (Dias et al., 2018). The information score of the expert e_i for the variable j is

$$I_j(e_i) = \sum_{k=0}^{t-1} p_{k+1} \ln\left(\frac{p_{i+1}}{(x_{e_i k+1} - x_{e_i k})/\Delta d}\right).$$

The average information score of the expert e_i is the average of all information scores of seed and target variables q_{all} the expert has answered

$$I(e_i) = \frac{\sum_{j=1}^{q_{all}} I_j(e_i)}{q_{all}}. \tag{9}$$

The experts are weighted according to the results of the calibration and information scores. The weights are used in order to construct probability distributions for the DM with weighted averages. The raw global weights w_i are defined as a product of the calibration score (7), the average information score (9) and the value of the indicator function (8)

$$w_i = \text{Cal}(e_i) \cdot I(e_i) \cdot 1_\alpha(\text{Cal}(e_i)). \tag{10}$$

In addition, we obtain the global weights by normalizing the raw weights

$$w_i' = \frac{w_i}{\sum_{i=1}^{m} w_i}. \tag{11}$$

Additionally, one can test the performance of the DM by calculating the DM calibration score. This is done by calculating a weighted average of the estimations on the seed variables and check how the true values place in them and calculate the calibration score (7). We usually want our DM to have a better calibration score than the experts, and this can be done by optimizing the weigths by increasing the threshold α until the DM calibration score maximizes.

We can now move on to eliciting marginal probabilities for continuous target variables. The idea is to construct a cumulative distribution function (CDF) from the estimations the expert has given. We elicit the 5th, 50th and 95th

quantiles of the target variable, but in some cases also the 25th and 75th quantiles are included in order to obtain better accuracy, for this, the calibration and information score must be calculated also with the additional quantiles. The minimum and maximum values of the target variable are again determined by the analyst or the experts. After all experts have answered the questions, we interpolate the answers of every expert e_i with either the linear or the quadratic method, keeping in mind that interpolation can cause errors. The DM CDF is the weighted average of all experts CDFs

$$P_{DM}(X \leq x) = \sum_{i=1}^{m} w_i' P_{e_i}(X \leq x),$$

where w_i' is the normalized weight in (11) and $P_{e_i}(X \leq x)$ is the interpolated cumulative probability of the expert e_i.

In PCIA, we want to have intervals for the marginal probabilities, and therefore we calculate a confidence interval (CI) for the DM probability. Hence, optimizing the weights for better DM quality is not necessary, because we are not interested in exact values. In order to calculate the confidence interval at a certain significance level β, we decided to use the bootstrap confidence interval 1, discussed in Gatz and Smith (1995), because it only assumes the data is independent and identically distributed, and parametric methods for CI calculations are not proper, because the number of experts in set E is usually not large and we can not be sure what the distributions for the estimations are. Errors in the bootstrap CI may also occur, but they can be fixed by increasing the iteration number.

Algorithm 1: Bootstrap confidence interval

Data $\bar{x} = (x_1, x_2, \ldots, x_n)$, iterations k, significance level β ;
for *i = 1:k* **do**
  1. Select n datapoints randomly from data $\bar{x}$ with replacement to create a new sample $x^* = (x_1^*, x_2^*, \ldots, x_n^*)$;
  2. Calculate a new weighted average from the new sample x^*;
end
3. Order estimated weighted averages from smallest to largest;
4. Calculate a 100(1-β)% level confidence interval by choosing $[k \times (\beta/2)]$ ordered estimate as lower endpoint and $[k \times (1 - \beta/2)]$ as the upper endpoint.;
return *Confidence interval*

To conclude, the probability intervals for every outcome of the uncertainty factor are calculated from the cumulative distributions formed by the lower and upper bounds of the calculated confidence interval of the decision maker CDF.

4.3 Discrete probability distributions

Calibration for discrete variable elicitation is performed with the following methodology presented in Dias et al. (2018) and the IDEA protocol (Hanea et al., 2017). First, we start with generating a q_d number of seed variables, which are events with a certain probability of occurrence. Next, we define a number n_b of probability bins $b_j = (p_j, 1 - p_j)$, where p_j is the probability of occurrence. For example, $b_1 = (0.2, 0.8)$, $b_2 = (0.5, 0.5)$ and $b_3 = (0.7, 0.3)$. We ask our experts to assign seed variables to a corresponding bin, where they believe the probability of occurrence lies. Let n_j be the amount of seed variables assigned to a bin b_j by an expert, and let r_j be the proportion of these seed variables that actually occur. The relative information between r_j and p_j is calculated with $I(r_j, p_j) = r_j \ln(\frac{r_j}{p_j}) + (1 - r_j)\ln(\frac{1-r_j}{1-p_j})$. As a result, for n_j independent seed variables, with the occurrence probability p_j, a measure of $2n_j I(r_j, p_j)$ is asymptotically Chi-squared distributed with one degree of freedom (Dias et al., 2018). So for n_b bins, we calculate the calibration score as follows

$$\text{Cal}(e_i) = 1 - \chi^2_{n_b}\{X < \sum_{j=1}^{n_b} 2n_j I(r_j, p_j)\}.$$

The average informativeness of the estimations for the seed variables is calculated by comparing how many times the seed variables have been placed to the uniform bin $b = (0.5, 0.5)$ (Dias et al., 2018)

$$I_s(e_i) = \frac{1}{q_d} \sum_{j=1}^{n_b} n_j I(p_j, 0.5).$$

The information score for q_t discrete target variables are calculated in the following way. Let the probability estimations for the uncertainty factor Y_l with K_l number of outcomes be $P_{Y_l} = [P^1_{Y_l}, P^2_{Y_l}, ..., P^{K_l}_{Y_l}]$. Let the uniform distribution be $u_{Y_l} = 1/K_l$. The information score for one target variable $I_{e_i}(u_{Y_l}, P_{Y_l})$ is calculated with the divergence (6). Hence, the average information score for the target variables is calculated as $I_t(e_i) = \frac{1}{q_t} \sum_{l=1}^{q_t} I_{e_i}(u_{Y_l}, P_{Y_l})$. If the DM wants to include the informativeness of the target variables in the weighting, then the average information score is the mean of the seed and target variable information scores $I(e_i) = (I_s(e_i) + I_t(e_i))/2$, otherwise we use only the seed variable calibration. The global raw weights are $g_i = \text{Cal}(e_i) \cdot I(e_i) \cdot 1_\alpha(\text{Cal}(e_i))$, and the normalized weights g'_i are calculated according to (11).

Now, we move on to the reference lottery method, where the idea is create two options, shown in Figure 4. The first option is the lottery M and the other is the reference lottery R. The expert is asked to choose between the two lotteries and the assigned probability pr is changed depending on which lottery the

expert preferred. In conclusion, the estimated probability $P(A)$ is determined when the expert is indifferent between the two lotteries $M \sim R$. More formally

$$W \cdot P(A) + L \cdot (1 - P(A)) = W \cdot pr + L \cdot (1 - pr),$$

where W is the price or severity of event A occurring and L is the price or severity of event A not occurring. The process starts with estimating the

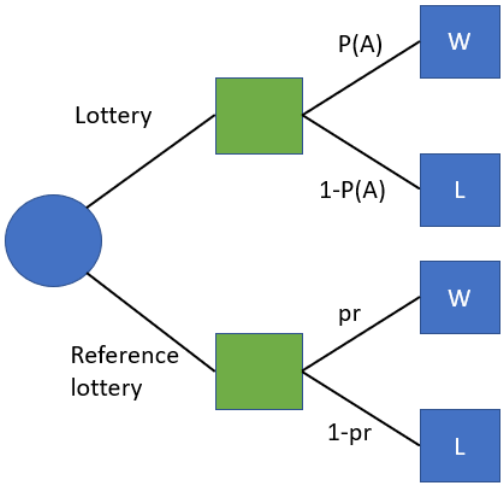

Figure 4: Reference lottery.

probability $P_{e_i}(A_j^1)$ of the first outcome A_j^1 of the uncertainty factor Y_j. The next step is to define the initial probability pr_1, which is an enlightened guess for the probability of the outcome A_j^1. Next, we start eliciting the preference of the expert between the two lotteries for l rounds. The assigned probability pr_l is changed according to the preference of the expert and the bound for the pr probability also changes. If lottery M is preferred, then $pr_{l+1} > pr_l$ and $pr_{\min} = pr_l$. If reference lottery R is preferred, then $pr_{l+1} < pr_l$ and $pr_{\max} = pr_l$. The new probability pr_{l+1} is chosen randomly between $[pr_{\min}, pr_{\max}]$. The elicitation for the probability of the outcome A_j^1 is ended when the expert e_i is indifferent between the lottery and reference lottery. The estimated probability $P_{e_i}(A_j^1)$ is returned.

The probabilities for the rest of the outcomes are estimated with the same procedure. However, the probability for the next outcome A_j^{k+1} is between $[0, 1 - \sum_{x=1}^{k} P_{e_i}(A_j^x)]$, and therefore the first value pr_1 must be between that interval. This process is performed $K_j - 1$ times, where K_j is the number of outcomes of the uncertainty factor Y_j. The estimate for the last probability $P_{e_i}(A_j^{K_j})$ is calculated as the complement with respect to the other outcome probabilities: $P_{e_i}(A_j^{K_j}) = 1 - \sum_{x=1}^{K_j - 1} P_{e_i}(A_j^x)$. The DM probability is calculated

with a weighted average

$$P_{DM}(A_j^k) = \sum_{i=1}^{m} g_i' P_{e_i}(A_j^k),$$

where g_i' is the normalized weight. The confidence interval for the DM probability is calculated by using the bootstrapping algorithm 1. However, the bootstrap algorithm may give probability intervals that do not sum up to one, and therefore the lower and upper end points are normalized in a way that the results are proper probability distributions.

4.4 Cross-impact judgements

In order to acquire judgements efficiently on cross-impact multipliers, we create a consistency table, which indicates the level of consistency of two outcomes with a strict qualitative description. We use this method, because numerical descriptions might be difficult to understand if the concept of cross-impact multipliers is not familiar. These qualitative consistency tables have been used in Weimer-Jehle (2006) and Seeve (2018). However, we must describe the

Table 2: Consistency descriptions to pairs of outcomes A and B.

Level c	**Description**	**Interval** Q_c
3	Strongly increasing. The occurrence of outcome A strongly increases the probability of outcome B to occur.	$\delta^2 - \delta^z$
2	Increasing. The occurrence of outcome A increases the probability of outcome B to occur.	$\delta^1 - \delta^2$
1	Slightly increasing. The occurrence of outcome A slightly increases the probability of outcome B to occur.	$\delta^0 - \delta^1$
0	Independent. The outcomes occur independently and have no direct relation.	$\delta^0 \pm \delta/10$
-1	Slightly reducing. The occurrence of outcome A slightly reduces the probability of outcome B to occur.	$\delta^{-1} - \delta^0$
-2	Reducing. The occurrence of outcome A moderately reduces the probability of outcome B to occur.	$\delta^{-2} - \delta^{-1}$
-3	Strongly reducing. The occurrence of outcome A strongly reduces the probability of outcome B to occur.	$\delta^{-z} - \delta^{-2}$

consistencies carefully, because bad qualitative descriptions can be misleading and they might not describe the relationship of the pair of outcomes in a proper

way. As mentioned before, the cross-impact multipliers are non-negative values, and therefore in order to convert the consistency levels c from Table 2 into a form that can be used in calculations, we use a parameter $\delta \in \mathbb{R}_{\geq 1}$, and we can define a specific level of consistency as cross-impact multiplier interval Q_c shown in Table 2. The parameter z is an upper limit to the power of the parameter δ, if one is deemed necessary.

To start the calibration, a pool of seed variables are generated. Each seed variable describes the cross-impact multiplier of two outcomes according to the definition (1). The experts are shown the Table 2, which contains descriptions of the consistency levels and their numerical intervals. Next, the experts are asked to assign levels $c \in [-3, 3]$ to represent the consistency of the outcomes in the seed variables at a somewhat certain confidence. However, it is challenging to assign a strict level to represent a consistency, and therefore the experts can give judgements such as "the level lies between 0 and 2", which indicates that the expert believes the true cross-impact multiplier lies in the interval $[Q_0^{\text{low}} - Q_2^{\text{upp}}]$. Thus, estimating an interval $[-3, 3]$ is the most uninformative.

We create three ($t = 3$) intervals for the expected proportion $p = [p_1, p_2, p_3]$. Depending on the number and difficulty of the seed variables, the expected proportion can be changed. Generally, we want our experts to estimate the cross-impacts multipliers correctly at a chance of $60 - 90\%$, meaning the minimum expected proportion is $[0.2, 0.6, 0.2]$ and the maximum $[0.05, 0.9, 0.05]$. From the results we can calculate the observed proportion r. The calibration score $\text{Cal}(e_i)$ is calculated in the same way as in (7), where we use the Chi-squared distribution with two degrees of freedom. The information score is calculated for the seed and target variables by scoring the width of the estimated consistency level interval $[c^{\text{low}}, c^{\text{upp}}]$ with a linear function:

$$I_j(e_i) = 1 - \frac{|c^{\text{upp}} - c^{\text{low}}|}{6}$$

The average information score is $I(e_i) = \frac{1}{q_{all}} \sum_{j=1}^{q_{all}} I_j(e_i)$, where q_{all} is the number of seed and target variables. The global raw weights are calculated as in (10), i.e. $u_i = \text{Cal}(e_i) \cdot I_{ci}(e_i) \cdot 1_\alpha(\text{Cal}(e_i))$, and they are normalized as in (11).

The cross-impact multiplier elicitation starts by dividing every pair of uncertainty factors (Y_k, Y_l) into sub-matrices $C_{kl} \in \text{CIM}$. These sub-matrices are presented in the elicitation sheet, where the experts can manually give consistency judgements as presented in Figure 5 using the Table 2 as a reference. The sub-matrices that expert e_i has constructed are compiled into one CIM, which is denoted as C_{e_i}. During the process, the following instructions are key elements in order to minimize the workload of the experts and to obtain better judgements:

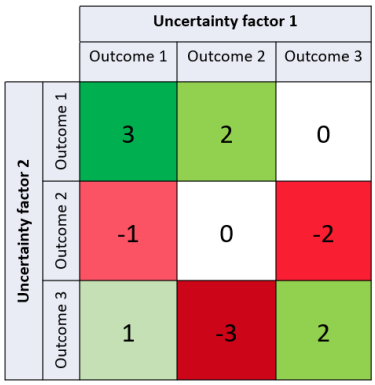

Figure 5: Example of cross-impacts judgements to a pair of uncertainty factors.

- If the number of cross-impacts multipliers is very large, then a certain number of most significant or potentially risky combinations of outcomes are selected, and the experts will assess only these.
- The experts are allowed to not give judgements on consistencies, where they feel very uncertain. The DM and the analyst will provide the proper solution. Intervals for consistency levels are allowed.
- The facilitator must guide the experts in giving mathematically realistic solutions in a way that the judgements will not be biased.

From the matrices C_{e_i}, we can calculate the lower and upper bounds for the DM cross-impact multipliers $[C_{\mathrm{DM}}^{\mathrm{low}} - C_{\mathrm{DM}}^{\mathrm{upp}}]$. Firstly, we convert the estimated consistencies c in the matrix C_{e_i} into intervals Q_c, and calculate the interval mean. Secondly, the confidence intervals are calculated for the weighted mean $C_{\mathrm{DM}} = \sum_{i=1}^{m} u_i' C_{e_i}$ using the bootstrap algorithm 1 at a 99% confidence level for more uncertainty. In situations, where all experts give the same level of consistency to a pair, we set the DM bounds to equal the interval of the specific level c according to Table 2. Finally, we remove independence assumptions using a parameter $\xi \in \mathbb{R}_{>0}$ in the following way:

- If $\delta^{-1/2} < C_{\mathrm{DM}}^{\mathrm{low}} \leq 1$ and $C_{\mathrm{DM}}^{\mathrm{upp}} > \delta$, then set $C_{\mathrm{DM}}^{\mathrm{low}} = 1 + \xi$
- If $1 \leq C_{\mathrm{DM}}^{\mathrm{upp}} < \delta^{1/2}$ and $C_{\mathrm{DM}}^{\mathrm{low}} < \delta^{-1}$, then set $C_{\mathrm{DM}}^{\mathrm{upp}} = 1 - \xi$.

Now, we proceed to test feasibility of the solution by running the optimization problem. Obtaining a non-feasible solution indicates that the cross-impact judgements are mathematically impossible, and therefore procedures for investigating what caused the non-feasibility must take place. To have a clue, where to begin investigating, we can run the optimization problem separately for every sub-matrix, and see if they get a feasible solution. This is not proven to be a correct method, but it gives a direction, where the problem might be.

By identifying the sub-matrices, where the incorrect intervals for cross-impacts may lie, we can try to widen the intervals for λ percents iteratively until they reach determined limits to the lower and upper bounds (lb, ub). The correction algorithm 2 aims to shift the cross-impact multipliers closer to the value one. The algorithm can be found in the appendix A. After this procedure, we can again check if removing independence assumptions can be done. The final option if the methods above are not working is to perform the elicitation for the particular sub-matrices again. Moreover, we must observe if the intervals for the cross-impact multipliers are too wide, because even if it may lead to feasible solutions, the informativeness of the cross-impact multipliers are not appropriate. For example, if the interval is $C_{ij}^{kl} \in [0.45 - 2.33]$, it indicates that the consistency of the outcomes A_i^k and A_j^l can be anything. Hence, considering a new cross-impact elicitation for this is justified.

5 Experimental case study

5.1 Problem setup

The developed elicitation process is used in an experimental case study, where the object is to test the efficiency of the process, analyse the quality of the judgements and to evaluate the total risk of the system. The main focus was on evaluating the risk of graduation delay from scenarios caused by the COVID-19 crisis. The target group was bachelor students whose graduation is planned for the academic year of 2020-2021. The experts in this study are five students, whose identities are kept anonymous, however, their backgrounds are highlighted in Table 3. The study focuses on four different uncertainty

Table 3: Backgrounds of the selected experts.

Expert 1	Computer Science
Expert 2	Mechanical and Structural Engineering
Expert 3	Engineering Physics and Mathematics
Expert 4	Industrial Engineering and Management
Expert 5	Chemical Engineering

factors that can affect the graduation of the bachelor student. The uncertainty factors and their outcomes are illustrated in Table 4, and the explanations for the factors are explained below.

The first uncertainty factor is the COVID-19 situation in Finland during the academic year of 2020-2021. The focus is on estimating the probabilities for

Table 4: An example of a scenario colored in blue. Top row contains the uncertainty factors and the columns below contain their outcomes.

COVID-19 situation (Discrete)	**Course arrangements (Continuous)**	**Student performance (Continuous)**	**Financial situation of a student (Continuous)**
Stable	Normal: 0-10% courses online.	Low: $<$ -5 credit difference	Impaired: < -50 euros monthly
Moderate	Mixed: 11-80% courses online	Normal: ± 5 credit difference	Normal: ± 50 euros monthly
Critical	Remote: 81-100% courses online.	High: $>$ 5 credit difference	Improved: > 50 euros monthly

three different discrete outcomes that include the virus spreading rate and restriction protocols with respect to the situation of July 2020 in Finland. This is a parameter that defines the overall behaviour of the society, and therefore can correlate highly with other factors.

- **Stable**: The number of infections will remain at or below the level of July 2020 and the phasing out of restrictions will continue, with restrictions completely lifted by the end of 2020.
- **Moderate**: Infections will rise again to the level of March 2020, and restrictions will be reintroduced to the same extent as in spring 2020.
- **Critical**: The number of infections will rise to an unprecedented level in Finland, and much higher restrictions will be introduced than in spring 2020.

The second uncertainty factor is course arrangements. This factor estimates the percentage of the courses organized online in Aalto University during the academic year of 2020-2021, and it may correlate highly with the COVID-19 situation, because universities change their operative strategies often according to the recommendations of the government and health authorities. Furthermore, this factor can affect the students studying methods and motivation in a negative or a positive way, depending on the preferences of the student.

The third uncertainty factor is student performance. This factor targets in measuring the difference of credits obtained in the year of studies compared to the credit goal of the individual student. The difference to the goal is measured, because bachelor students planned to graduate within 2020-2021, may have different numbers of credits left to obtain in order to complete the requirements of the degree.

The fourth uncertainty factor is the financial situation of a student. This factors aims in measuring the difference in the students monthly operating assets, that is, the amount of money the student uses monthly for basic needs. This factor is interesting, because there might be significant differences due to the fact that many students lost their summer job due to the COVID-19 crisis.

The risk analysis for the system is done by calculating the expected utility. We have estimated every outcome an expected utility $U \in [-1, 1]$, which indicates how the outcome speeds up the graduation of a bachelor student. Negative values indicate delay and positive values indicate acceleration in the graduation. Value zero means the outcome has no effect. The scenario specific expected utilities are calculated as averages of the outcome utilities, and they can be interpreted as time utilities. We can specify the utility $U = -1$ to indicate five period delay (one academic year) and $U = 1$ to indicate five period speed up. Using this information we can say that a utility $U = -0.2$ would mean a one period delay. The outcome specific expected utilities are shown in Table A1 in the appendix.

We used the following expected proportions and parameters in the process. For continuous variable calibration, we used the expected proportion of $p = [0.05, 0.45, 0.45, 0.05]$ for the 5th, 50th and 95th quantiles. For cross-impact multiplier calibration we used an expected proportion of $p = [0.2, 0.6, 0.2]$, which means we looked for 60% of correct estimations and 20% under- and overshoots were allowed. This is important to keep in mind, because the experts might give under/overconfident judgements. The threshold level for calibration scores was set to $\alpha = 0.01$, the delta parameter for the cross-impact multipliers was set to $\delta = 1.3$, the maximum power for the delta parameter was $z = 4$, the independence removing parameter was $\xi = 0.05$ and the correction parameter $\lambda = 0.08$. The limits for the correction algorithm 2 are presented in Table A2.

Seed variables were formulated in a way that they represent the target variables as much as possible, and the information for them was acquired from the following websites: Finnish Institute for Health and Welfare (Terveyden ja hyvinvoinnin laitos, 2020), Statistics Finland (Tilastokeskus, 2020), The Social Insurance Institution of Finland (Kela, 2020), Aalto University (2018), Aalto University (2019) and Aalto University Learning Centre (2020). Both probability and cross-impact multiplier calibrations use eight seed variables. We decided, to calculate the probability weights only according to continuous variable calibration, because the problem has few uncertainty factors and only one of them is discrete. The elicitation sheet and the judgements of the experts are presented in appendix A.

5.2 Process analysis

The process was performed separately for every expert using remote connections, following the steps explained in 4.1. We performed one main round for the elicitation and one fast round, where the experts gave some refining judgements on the target variables. The elicitation sheet covered the calibration phase in parts one and two, and the target variable elicitation phase in parts three and four. For every expert, we presented the main idea of the process and carefully explained the instructions. Overall, the process worked well, and most of the confusion occurred only at the beginning, as the experts were getting used to the types of questions and developing a routine on how to answer them. Manually giving judgements for the various questions was considered to be easy among the experts, however, constant contemplation and giving judgments was becoming more tough when time went by, so breaks had to be taken.

The times taken by the experts in the first round, refining round and overall remained within tolerable limits and are presented in Table 5. The average overall time taken by the experts was approximately 100 minutes. Keeping in mind that our study had only 16 seed variables in total and 66 target variables, the time taken by one expert on average was quite good. Performing a larger study containing ten times more seed and target variables will probably cost an entire day of work (8-10 hours) for one expert.

Table 5: Times taken by experts at each round.

Expert	1	2	3	4	5
Time of round 1	131 min	53 min	66 min	134 min	33 min
Time of round 2	10 min	15 min	25 min	15 min	10 min
Overall time	141 min	68 min	91 min	149 min	43 min

In the calibration phase, the experts were not told that they are answering calibration questions, but they understood by themselves that the questions have a known answer, and that the results might qualify them in some way. This indicates that, in practice, it is not always feasible to keep the calibration part unknown from the experts. However, for future studies, the analyst could merge seed and target questions together into the same phase.

The target variable elicitation phase was considered to be the easiest among the experts, although, a somewhat large amount of thinking was required in order to understand pairwise relationships better in the cross-impact elicitation part. In probability elicitation, the reference lottery method used for the COVID-19 uncertainty factor was considered to be hard to understand at the beginning, and therefore took more time than the other methods. From this it can be deduced that eliciting continuous variables is more efficient than discrete variables.

Additionally, for continuous variables the workload for eliciting probabilities is smaller than in discrete variables. Consider two uncertainty factors (continuous and discrete) with ten outcomes each. For the continuous uncertainty factor we are able to get probabilities for every outcome with just one lower bound, median and upper bound estimation. For the discrete one, the expert needs to estimate nine different probabilities and the last one is calculated as the complement. Giving cross-impact judgements with a qualitative scale from -3 to 3 was considered to be an easy an fast way in giving the pairs of outcomes a representation of their relationship.

One refining round of elicitation was performed, because the first judgements had some inconsistencies and big differences, and the Delphi method (Dalkey and Helmer, 1963) aims in pushing the judgements of the experts in the same direction. The experts were not shown the judgements of the others, but they were told the direction and approximate magnitudes of the judgements. This resulted in some new and more consistent results. If the study would have been larger, more refining rounds for target variables could have been performed.

5.3 Results

The following data was obtained from the elicitation process. The calibration scores, information scores and weights for probability and cross-impact judgements are presented in Figure 6, which are color scaled from red (worst) to green (best).

Probability judgements

Expert	1	2	3	4	5
Calibration	0.1850	0.0426	0.5405	0.0877	0.5405
Informativeness	0.5202	1.0699	0.1395	0.6605	0.4305
Weight	0.1896	0.0897	0.1485	0.1140	0.4582

Cross-impact multiplier judgements

Expert	1	2	3	4	5
Calibration	0.1680	0.0310	0.5030	0.5030	0.3980
Informativeness	0.9792	0.9583	0.9375	1.0000	0.8958
Weight	0.1093	0.0195	0.3069	0.3376	0.2266

Figure 6: Calibration scores, information scores and normalized weights of probability and cross-impact judgements.

It is useful to keep in mind that we used only eight seed variables for each target

variable type, which means the observed proportions of the experts always differ from the expected proportions, meaning, the calibration scores may look bad, but actually they are not. For example, experts three and five gave very good probability judgements on the seed variables, but got only a score of ≈ 0.54. Overall the experts performed well in the calibration phase. They performed better in the probability calibration compared to the cross-impact calibration, where we used a more flexible expected proportion. If we would have used the most strict expected proportion $p = [0.05, 0.9, 0.05]$ for the cross-impacts, only two experts would have qualified from the calibration and their calibration score would not have been the best. These calibration results indicate that the probability judgements of the experts are quite reliable, but the cross-impact judgements must be critically evaluated as under/overshoots can occur in target variables.

Final refined judgements for the target variables regarding the probabilities are visualized in Figure 7. The DM probabilities with the 95% confidence intervals are illustrated in Figure 8, and the numerical marginal probabilities for the outcomes are shown in Table 6. In the light of the calibration scores of the experts in Figure 6, we can say that using 95% confidence intervals for the probabilities give reliable and useful predictions for the future.

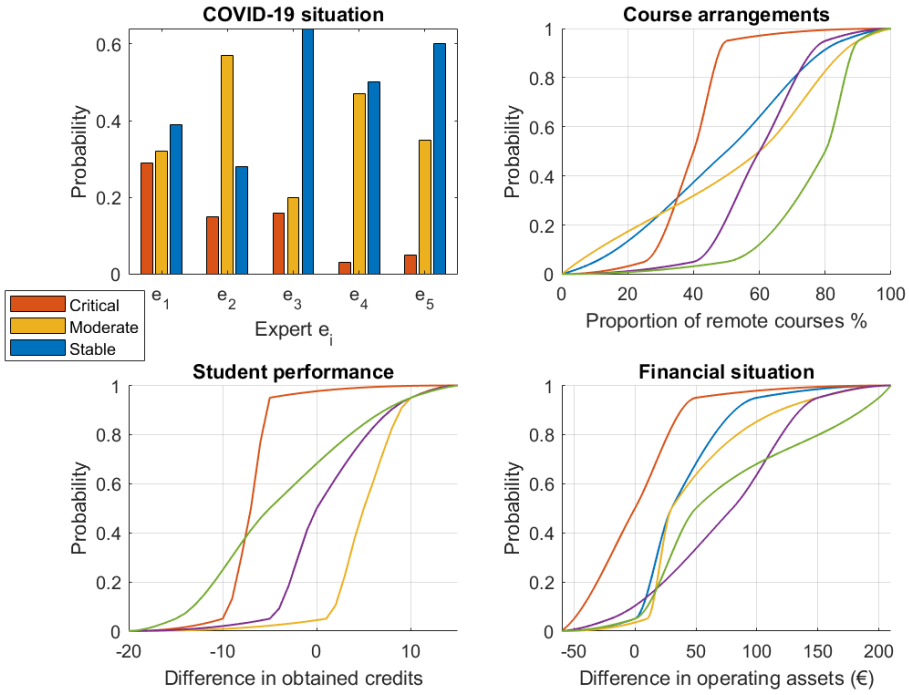

Figure 7: Probability judgements.

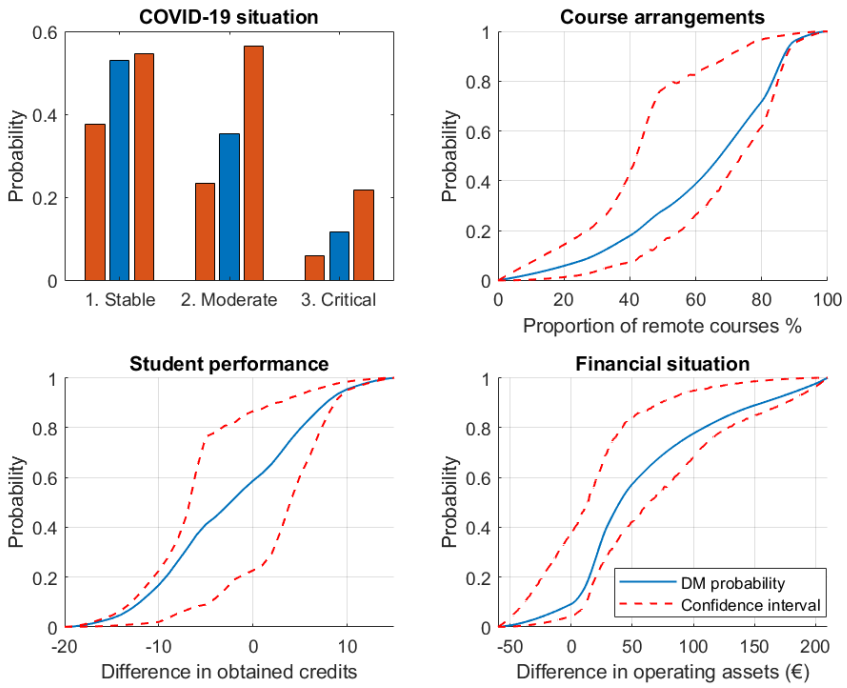

Figure 8: Constructed DM probabilities and 95% confidence intervals.

Table 6: Marginal probabilities

Uncertainty factor	Outcome	Marginal probability
COVID-19	Stable	$[0.3744 - 0.5504]$
	Moderate	$[0.2319 - 0.5648]$
	Critical	$[0.0608 - 0.2177]$
Course arrangements	Normal	$[0.0041 - 0.0798]$
	Mixed	$[0.6116 - 0.8857]$
	Remote	$[0.0345 - 0.3843]$
Student performance	Low	$[0.0770 - 0.6364]$
	Normal	$[0.2966 - 0.5302]$
	High	$[0.0669 - 0.3928]$
Financial situation	Impaired	$[0.0009 - 0.0302]$
	Normal	$[0.4177 - 0.8036]$
	Improved	$[0.1661 - 0.5815]$

In the cross-impact multiplier elicitation, the experts gave judgements to every pair of outcomes, with some refining judgements in the second round. Every pair of uncertainty factors obtained dependent judgements except the pairs (Course arrangements, Financial situation) and (Student performance, Financial situation), which were deemed independent. In the result analysis, we did not obtain a feasible solution for the optimization problem immediately, so we applied independence removing and we widened the cross-impacts of the sub-matrices (COVID-19, Financial situation) and (Course arrangements, Student performance).

Figure 9 shows the dependent uncertainty factors and their cross-impact multipliers with 99% confidence intervals, where green cells indicate an increasing effect in the cross-impact multiplier and red cells indicate a decreasing effect. Due to the calibration scores, the cross-impact multipliers are still not the most reliable even with 99% confidence intervals. Some cross-impacts have very wide independent intervals, for example, COVID-19: critical and course arrangements: mixed, where the interval is $[0.2959 - 1.3956]$. This means we can not really tell anything about their relationship. The experts, however, where satisfied with their own judgements and as decision makers we can only give controlled feedback, but not force them to change their thinking.

		Course arrangements		
		Normal	Mixed	Remote
COVID-19	Stable	1.2150 – 1.4950	1.0500 – 2.6170	0.3448 – 0.8019
	Moderate	0.2959 – 0.6805	0.6805 – 1.3940	1.3353 – 2.8450
	Critical	0.3501 – 0.5917	0.2959 – 1.3956	1.6900 – 2.8561
		Student performance		
		Low	Normal	High
COVID-19	Stable	0.7074– 0.9500	1.0500 – 2.6101	1.0500 – 1.4320
	Moderate	0.9002 – 1.4555	0.8196 – 1.0000	0.6805 – 1.2150
	Critical	1.0500 – 2.6600	0.7017 – 1.1647	0.2959 – 1.2150
		Financial situation		
		Impaired	Normal	Improved
COVID-19	Stable	0.7498 – 1.1897	1.0500 – 2.4983	0.8635 – 1.4320
	Moderate	0.7089 – 0.9410	0.8196 – 1.1682	1.1197 – 1.6170
	Critical	0.5453 – 0.8320	0.7157 – 0.9770	1.3778 – 2.8812
		Student performance		
		Low	Normal	High
Course arrangements	Normal	0.6880 – 0.9880	1.0080 – 2.7195	0.6871 – 0.9880
	Mixed	1.0083 – 2.7309	0.8196 – 1.1841	1.0100 – 2.7386
	Remote	1.1664 – 2.7430	0.6677 – 0.8524	1.1664 – 2.7430

Figure 9: Confidence intervals at 99% for the cross-impact multipliers. Rest of the uncertainty factor pairs were estimated to be independent, i.e. $[\delta \pm \delta/10] = [0.8700 - 1.1300]$.

The expected risk was calculated by minimizing and maximizing the expected utility of the system. We obtained a lower bound $\text{Risk}^{\text{low}} = -0.23$ and an upper bound $\text{Risk}^{\text{upp}} = 0.16$, which would indicate that in the worst case, the graduation of a third year bachelor student will be delayed with one period, and in the best case, the graduation will speed up with almost one period. An interesting observation was that by running the optimization problem by setting every uncertainty factor independent from each other, we obtained a lower bound $\text{Risk}^{\text{low}}_{\text{ind}} = -0.27$ and an upper bound $\text{Risk}^{\text{upp}}_{\text{ind}} = 0.05$, which means that in the worst case the graduation would delay with approximately one and a half period and in the best case the student will approximately graduate on time. Basically, with the cross-impact judgements of the experts, we predict a larger variation in the expected utility, which yields a slightly better worst case and a better best case utility compared to the risk assessment with independent uncertainty factors. The risks are visualized in Figure 10.

The result for the risk assessment is not the most reliable as the calibration scores for the cross-impacts judgements were not the best. In addition, the expected utilities for the outcomes and scenarios were estimated subjectively without any judgements from experts, and therefore obtaining better utility judgements would yield a different result in the risk assessment. However, the marginal probabilities and cross-impact judgements estimate somewhat well the direction of what might happen in the future.

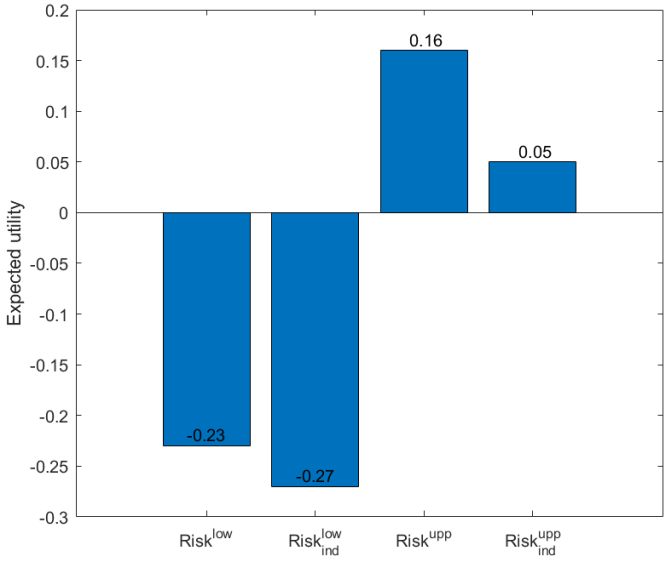

Figure 10: Lower and upper bounds of the dependent and independent risk assessment.

6 Conclusion

This thesis developed a process for eliciting expert judgements for probabilistic cross-impact assessment purposes, which aimed in making scenario based risk evaluations more efficient. The process was tested in an experimental case study, where we produced a risk assessment for a system with 81 potential scenarios. As a whole, the process worked as expected with the exception that it had to be performed remotely. The judgements for the marginal probabilities produced reliable data, but the cross-impact judgements have the potential to be under/overconfident, and therefore the results of the risk assessment may not be sufficient to make decisions.

According to the empiric results in the process analysis 5.2, we can conclude that the process is quite efficient overall in eliciting judgements for approximately 80 variables (seed+target). For larger scale studies it is necessary that the experts assemble in one place and use an allotted time for the process, because even if all experts would be capable to use an entire day for the process remotely, it becomes hard and time taking for the analyst and facilitator to use an entire day for just one expert at a time. It is important that the experts are not left without help, because also in our process the experts would have given illogical judgements. However, as we have learned in the current circumstances, working remotely is increasing fast. For future applications, it would be useful to turn this process into a web based software that can handle multiple experts giving judgements at the same time in different locations, independently, anonymously and with the assistance of a facilitator. With a web based tool, it would be possible to study implementing a Bayesian network in the process, which would update the prior distributions of the data by gathering real time information on events as time goes by. Additionally, in terms of workload, it would be useful to focus on continuous variables as much as possible.

The reliability of the produced data is strongly based on the calibration results of the experts. If we want our experts to be calibrated with better precision, we must generate more seed variables and acquire information on them. This is a laborious process for the analyst, but it would produce better results as we get more accurate observed proportions for the experts, which can be compared to the expected proportions. Additionally, the performance of the experts can be improved by giving controlled feedback after the elicitation. The number of experts participating in the elicitation is also important, because it would be useful to have many opinions. A panel of five experts is quite small for giving judgements on high risk systems. The more experts the better, but it would also mean that the judgements can have a lot of variance, and therefore multiple refining rounds for the target variables are likely to be needed.

Most difficulties in eliciting and analysing data occurred with the cross-impact judgements. Eliciting cross-impacts judgements requires the facilitator to observe and point out possible mathematical impossibilities, i.e. non-feasible solutions. This is quite laborious for the facilitator and it might lead to biased judgements. Furthermore, obtaining a non-feasible solution requires examining what caused it. Our method to test feasibility one pair of uncertainty factors at a time performed well, but our data was quite small. For larger data sets this might not be an appropriate method. For a potential web based software, it is worth considering testing the feasibility automatically as the experts give their judgements on cross-impacts.

This thesis does not examine or take position on how the expected consequences or utilities for risk assessment purposes are elicited. Thus, for future research one can implement the elicitation for consequences and utilities into the process presented here.

We are grateful for all the experts who volunteered their time to help with the study. It is worth mentioning that the experts are only third year students with no work experience in epidemiology, economics or university management, and in this sense they actually gave very good judgements. In addition, none of the experts are considered at any point bad or good, only their judgements are qualified in a mathematical way.

References

Aalto University. Key figures of 2018 and reports. 2018. URL https://www.aalto.fi/en/aalto-university/key-figures-of-2018-and-reports.

Aalto University. Key figures of 2019 and reports. 2019. URL https://www.aalto.fi/en/aalto-university/key-figures-of-2019-and-reports.

Aalto University Learning Centre. Bachelor's theses. 2020. URL https://aaltodoc.aalto.fi/handle/123456789/1.

J. R. Avella. Delphi panels: Research design, procedures, advantages, and challenges. *International Journal of Doctoral Studies*, 11(1):305–321, 2016.

V. A. Bañuls and M. Turoff. Scenario construction via Delphi and cross-impact analysis. *Technological Forecasting and Social Change*, 78(9):1579–1602, 2011.

N. Dalkey and O. Helmer. An experimental application of the Delphi method to the use of experts. *Management science*, 9(3):458–467, 1963.

L. C. Dias, A. Morton, and J. Quigley. *Elicitation: The Science and Art of Structuring Judgement*. Springer, New York, 2018.

C. A. Doswell III. Weather forecasting by humans—heuristics and decision making. *Weather and Forecasting*, 19(6):1115–1126, 2004.

D. F. Gatz and L Smith. The standard error of a weighted mean concentration—i. bootstrapping vs other methods. *Atmospheric Environment*, 29(11): 1185–1193, 1995.

A. M. Hanea, M. F. McBride, M. A. Burgman, B. C. Wintle, F. Fidler, L. Flander, C. R. Twardy, B. Manning, and S. Mascaro. Investigate discuss estimate aggregate for structured expert judgement. *International Journal of Forecasting*, 33(1):267–279, 2017.

O. Helmer. Problems in futures research: Delphi and causal cross-impact analysis. *Futures*, 9(1):17–31, 1977.

G. H. Jeong and S. H. Kim. A qualitative cross-impact approach to find the key technology. *Technological Forecasting and Social Change*, 55(3):203–214, 1997.

J. Kangaspunta and A. Salo. Expert judgments in the cost-effectiveness analysis of resource allocations: A case study in military planning. *OR Spectrum*, 36 (1):161–185, 2014.

R. L. Keeney and D. von Winterfeldt. On the uses of expert judgment on complex technical problems. *IEEE Transactions on Engineering Management*, 36(2):83–86, 1989.

Kela. Etuuksien saajat. 2020. URL `https://koronamittarit.kela.fi/etuuksien_saajat.html`.

M Kynn. The 'heuristics and biases' bias in expert elicitation. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 171(1):239–264, 2008.

T. G. Martin, M. A. Burgman, F. Fidler, P. M. Kuhnert, S. Low-Choy, M. McBride, and K. Mengersen. Eliciting expert knowledge in conservation science. *Conservation Biology*, 26(1):29–38, 2012.

J. E. Oakley. Eliciting univariate probability distributions. *Rethinking Risk Measurements and Reporting*, 1, 2010.

J. E. Oakley and A. O'Hagan. Shelf: the sheffield elicitation framework (version 4). *School of Mathematics and Statistics, University of Sheffield*, 2019. URL `http://www.tonyohagan.co.uk/shelf/`.

A. Salo, E. Tosoni, J. Roponen, and D. W. Bunn. Cross-impact analysis for probabilistic risk assesment. *manuscript*, 2020.

T. Seeve. A structured method for identifying and visualizing scenarios. Master's thesis, Aalto University, School of Science, Espoo, 2018.

Terveyden ja hyvinvoinnin laitos. Koronaviruksen seuranta. 2020. URL `https://thl.fi/fi/web/infektiotaudit-ja-rokotukset/ajankohtaista/ajankohtaista-koronaviruksesta-covid-19/tilannekatsaus-koronaviruksesta/koronaviruksen-seuranta`.

Tilastokeskus. Suomen virallinen tilasto (SVT): Työvoimatutkimus [verkko-julkaisu] Helsinki: viitattu [27.08.2020]. Kesäkuu 2020. ISSN 1798-7830. URL `https://www.stat.fi/til/tyti/2020/06/tyti_2020_06_2020-07-21_tie_001_fi.html`.

E. Tosoni, A. Salo, J. Govaerts, and E. Zio. Comprehensiveness of scenarios in the safety assessment of nuclear waste repositories. *Reliability Engineering & System Safety*, 188:561–573, 2019.

J. von Neumann and O. Morgenstern. *Theory of games and economic behavior*. Princeton university press, 1953.

W. Weimer-Jehle. Cross-impact balances: A system-theoretical approach to cross-impact analysis. *Technological Forecasting and Social Change*, 73(4): 334–361, 2006.

A Appendix

Algorithm 2: Cross-impact correction

Set maximum correction rounds Rd. Set λ percentage increase. Set sub-matrix specific lower/upper bounds (lb_{ij}, ub_{ij});
for r = *1:Rd* **do**
  Test feasibility for every sub-matrix C_{ij} of the CIM;
  if *all sub-matrices* C_{ij} *feasible* **then**
    break;
  else
    for *every non-feasible* C_{ij} **do**
      for k = *1:size*$(C_{ij}, 1)$ **do**
        for l = *1:size*$(C_{ij}, 2)$ **do**
          if $C_{ij}(k,l)^{low} > ub_{ij}$ *or* $C_{ij}(k,l)^{upp} < lb_{ij}$ **then**
            $C_{ij}(k,l)^{low} = C_{ij}(k,l)^{low} \cdot (1 - \lambda/2)$;
            $C_{ij}(k,l)^{upp} = C_{ij}(k,l)^{upp} \cdot (1 + \lambda/2)$;
          end
        end
      end
    end
  end
end
Run optimization for whole system;

Elicitation sheet

Part 1. Answer the following questions. Give your estimates as 5^{th}, 95^{th} and 50^{th} quantiles. To be more specific, your answers are estimates of values so that the true value of the question has a 5%, 95% and 50% probability of being less or equal to the value you have estimated, i.e. the lower and upper bounds and the median.

Example: From a fleet of 1000 airplanes, how many planes will malfunction after 2000 hours of flight? ➔ 5^{th} = 2, 95^{th} = 20 and 50^{th} = 5.

1. Estimate how many confirmed infections of SARS-Cov-2 virus occurred in Finland during the week 32 (3.8.2020-9.8.2020).

5^{th}	95^{th}	50^{th}

2. Estimate how many COVID-19 test samples were performed in Finland during the week 32 (3.8.2020-9.8.2020). Give you answer in tens of thousands.

5^{th}	95^{th}	50^{th}

3. Estimate the overall incidence of SARS-Cov-2 virus cases in Finland. Give your answer in thousands or with more precision.

5^{th}	95^{th}	50^{th}

4. Estimate the percentage of students that have completed at least 55 credit points at Aalto University in the academic year 2019-2020.

5^{th}	95^{th}	50^{th}

5. Estimate the number of bachelor's degrees produced in Aalto University in academic year 2019-2020.

5^{th}	95^{th}	50^{th}

6. Estimate how many bachelor theses were published between January and July 2020 in Aalto University.

5^{th}	95^{th}	50^{th}

7. Estimate the difference in the number of working 15 to 24-year-olds in June 2020 compared to June 2019. Give your answers in tens of thousands or with more precision.

5^{th}	95^{th}	50^{th}

8. Estimate the number of new under 25-year-olds in June 2020 who will receive basic income support. Give your answer in thousands or with more precision.

5^{th}	95^{th}	50^{th}

Part 2. Consider the table below, which indicates the level of consistency of two outcomes, i.e., the cross-impact multiplier. The cross-impact multiplier is defined as the probability of outcome B given that outcome A occurred and divided by the probability of B, more formally C = P(B|A)/P(B).

Table 1.

Level	Description	C
3	Strongly increasing. The occurrence of outcome A strongly increases the probability of outcome B to occur.	1.69 - 4.82
2	Increasing. The occurrence of outcome A increases the probability of outcome B to occur.	1.3 - 1.69
1	Slightly increasing. The occurrence of outcome A slightly increases the probability of outcome B to occur.	1 – 1.3
0	Independent. The outcomes occur independently and have no direct relation.	0.87 - 1.13
-1	Slightly reducing. The occurrence of outcome A slightly reduces the probability of outcome B to occur.	0.77-1
-2	Reducing. The occurrence of outcome A moderately reduces the probability of outcome B to occur.	0.59-0.77
-3	Strongly reducing. The occurrence of outcome A strongly reduces the probability of outcome B to occur.	0.20-0.59

Below are presented eight (8) pairwise events. Select the level of cross-impact [-3,3] from the table above that corresponds to the pair of outcomes below. Give your answer as a specific level or as an interval. For example, the cross-impact lies at level 2 or between levels 2-3.

Outcome A	**Outcome B**	**Level**
COVID-19 situation in Finland in spring 2020.	A bachelor's thesis is published in Aalto University between January and June 2020.	
COVID-19 situation in Finland in spring 2020.	A 15-24-year-old working in June 2020 (Entire Finland).	
COVID-19 situation in Finland in spring 2020.	A working age (15-24) individual becomes an unemployed jobseeker in summer 2020 (Entire Finland).	
COVID-19 situation in Finland in spring 2020.	An individual is accepted to study in Aalto University. Consider the whole population of Finland.	
COVID-19 situation in Finland in spring 2020.	A course is arranged in the summer period of the year 2020 in Aalto University.	
A Student obtains a goal number of credit points in spring 2020.	A student's financial situation in summer 2020 stays unchanged compared to the situation in spring 2020.	
Restriction policies in Finland at the middle of June 2020 (15.6.-21.6.).	A COVID-19 test taken at the end of July 2020 (21.7.-26.7.) in Finland shows a positive result.	
Restriction policies in Finland at the end of June 2020 (22.6.-28.6.).	An individual person getting confirmedly infected by the SARS-Cov-2 virus during the week 32 (3.8.-9.8.) in Finland.	

Part 3. Answer the following questions in the same way as in Part 1 with your best capability. Give you answers as 5th, 95th and 50th quantiles. Assess every question independently.

1. Consider the courses in Aalto University. On average, what is the percentage of them that will be organized remotely during the academic year 2020-2021?

5th	95th	50th

2. On average, how much will the amount of obtained credits of a single student differ from his/her planned credit goal at the end of the academic year 2020-2021?

5th	95th	50th

3. How will the monthly operating assets (money available for basic needs) of a student change in the academic year of 2020-2021? Give your answers in euros.

5th	95th	50th

The following question requires assistance from the facilitator.

4. What is your estimate to the probability of Covid-19 to be in a stable / moderate / critical situation in the academic year 2020-2021?

Table 2.

COVID-19 outcomes in academic year 2020-2021
Stable: The number of infections will remain at or below the level of July 2020 and the phasing out of restrictions will continue, with restrictions completely lifted by the end of 2020.
Moderate: Infections will rise again to the level of March 2020, and restrictions will be reintroduced to the same extent as in spring 2020.
Critical: The number of infections will rise to an unprecedented level in Finland, and much higher restrictions will be introduced than in spring 2020.

Probability(Stable) =

Probability(Moderate) =

Probability(Critical) =

Part 4. Fill in the levels of consistencies ranging from -3 to 3 according to Table 1. The definitions to the COVID-19 outcomes were presented in Table 2. Furthermore, consider how the outcomes on the side A influence the outcomes on the side B. For example, if the outcome **COVID-19: Stable** occurs, how will it relate with the **remote course arrangements**.

Note: If you feel very uncertain about the consistency, leave the cell of the grid blank.

		Course arrangements (B)		
		Normal: 0-10% courses online	**Mixed: 11-80% courses online**	**Remote: 81-100% courses online**
COVID-19 (A)	**Stable**			
	Moderate			
	Critical			

		Student performance (B)		
		Low: < -5 credit difference	**Normal: ± 5 credit difference**	**High: > 5 credit difference**
COVID-19 (A)	**Stable**			
	Moderate			
	Critical			

		Financial situation of a student (B)		
		Impaired: Operating assets /month < -50 €	Normal: Operating assets /month ± 50 €	Improved: Operating assets /month > 50 €
COVID-19 (A)	Stable			
	Moderate			
	Critical			

		Student performance (B)		
		Low: < -5 credit difference	Normal: ± 5 credit difference	High: > 5 credit difference
Course arrangements (A)	Normal: 0-10% courses online			
	Mixed: 11-80% courses online			
	Remote: 81-100% courses online			

		Financial situation of a student (B)		
		Impaired: Operating assets /month < -50 €	Normal: Operating assets /month ± 50 €	Improved: Operating assets /month > 50 €
Course arrangements (A)	Normal: 0-10% courses online			
	Mixed: 11-80% courses online			
	Remote: 81-100% courses online			

		Financial situation of a student (B)		
		Impaired: Operating assets /month < -50 €	Normal: Operating assets /month ± 50 €	Improved: Operating assets /month > 50 €
Student performance (A)	Low: < 5 credit difference			
	Normal: ± 5 credit difference			
	High: > 5 credit difference			

Probability elicitation

Expert 1

	Min	5th	5th	95th	Max	Correct value
Seed 1	3	50	125	200	327	168
Seed 2	0	20000	80000	100000	215000	48969
Seed 3	100	3000	4500	5000	10900	7805
Seed 4	13.5	20	40	85	100	36 %
Seed 5	200	400	1100	1400	1600	1340
Seed 6	0	200	500	700	960	295
Seed 7	-109000	-20000	-10000	-3000	6000	-31000
Seed 8	0	2500	7500	9000	21900	10793
Target2	0	10	50	85	100	
Target 3	-20	-15	-5	10	15	
Target 4	-60	0	30	100	210	
Target 1	0.29	0.32	0.39		COVID-19: (Critical, Moderate, Stable)	

Expert 2

	Min	5th	5th	95th	Max
Seed 1	3	150	220	300	327
Seed 2	0	20000	30000	50000	215000
Seed 3	100	6999	7400	7500	10900
Seed 4	13.5	40	45	50	100
Seed 5	200	500	550	700	1600
Seed 6	0	700	750	900	960
Seed 7	-109000	-30000	-25000	-20000	6000
Seed 8	0	10000	18000	20000	21900
Target2	0	25	40	50	100
Target 3	-20	-10	-7	-5	15
Target 4	-60	-50	0	50	210
Target 1	0.15	0.57	0.28		COVID-19: (Critical, Moderate, Stable)

Expert 3

	Min	5th	5th	95th	Max
Seed 1	3	30	140	300	327
Seed 2	0	50000	100000	200000	215000
Seed 3	100	1000	3500	10000	10900
Seed 4	13.5	30	70	95	100
Seed 5	200	300	600	1000	1600
Seed 6	0	50	200	600	960
Seed 7	-109000	-100000	-30000	-10000	6000
Seed 8	0	1000	8000	20000	21900
Target2	0	5	60	90	100
Target 3	-20	1	5	10	15
Target 4	-60	10	30	150	210
Target 1	0.16	0.2	0.64		COVID-19: (Critical, Moderate, Stable)

Expert 4

	Min	5th	5th	95th	Max	
Seed 1	3	55	85	95	327	
Seed 2	0	20000	35000	50000	215000	
Seed 3	100	6000	7500	9000	10900	
Seed 4	13.5	25	40	60	100	
Seed 5	200	600	1000	1300	1600	
Seed 6	0	400	650	900	960	
Seed 7	-109000	-80000	-45000	-30000	6000	
Seed 8	0	2000	4000	6000	21900	
Target2	0	40	60	80	100	
Target 3	-20	-5	0	10	15	
Target 4	-60	-20	80	150	210	
Target 1	0.03	0.47	0.5		COVID-19: (Critical, Moderate, Stable)	

Expert 5

	Min	5th	5th	95th	Max	
Seed 1	3	70	140	210	327	
Seed 2	0	10000	40000	100000	215000	
Seed 3	100	6000	7000	10000	10900	
Seed 4	13.5	40	60	80	100	
Seed 5	200	800	1100	1500	1600	
Seed 6	0	300	600	900	960	
Seed 7	-109000	-40000	-20000	-10000	6000	
Seed 8	0	3000	7000	10000	21900	
Target2	0	50	80	90	100	
Target 3	-20	-15	-5	10	15	
Target 4	-60	0	50	200	210	
Target 1	0.05	0.35	0.6		COVID-19: (Critical, Moderate, Stable)	

Cross-impact elicitation

	Expert 1		**Expert 2**		**Expert 3**	
	low	**upp**	**low**	**upp**	**low**	**upp**
Seed 1	2	2	-2	-2	0	0
Seed 2	-3	-3	-2	-2	-1	-1
Seed 3	3	3	1	1	1	2
Seed 4	0	1	0	0	0	0
Seed 5	3	3	-2	-1	2	2
Seed 6	0	0	0	0	-1	0
Seed 7	-2	-2	0	1	-2	-2
Seed 8	-2	-2	0	0	-2	-1

	Expert 4		**Expert 5**		**Correct approximate answers**	
	low	**upp**	**low**	**upp**		
Seed 1	0	0	-1	0	Seed 1	-3
Seed 2	-1	-1	-2	-1	Seed 2	-1
Seed 3	1	1	1	2	Seed 3	2
Seed 4	0	0	-1	-1	Seed 4	0
Seed 5	3	3	0	0	Seed 5	3
Seed 6	0	0	0	0	Seed 6	0
Seed 7	-2	-2	-2	-1	Seed 7	-3
Seed 8	-1	-1	-2	-1	Seed 8	-3

Expert 1 Target variables

	Courses			Student perf			Financial		
	1	2	-2	-2	3	2	0	3	2
Covid-19	-3	-2	3	2	-1	-2	-1	-1	2
	-3	-3	3	3	-2	-3	-2	-2	3
				-2	3	-2	0	0	0
			Courses	3	1	3	0	0	0
				3	-2	3	0	0	0
							0	0	0
					Student perf		0	0	0
							0	0	0

Expert 2 Target variables

	Courses			Student perf			Financial		
	2	0	-2	-1	0	1	1	0	-1
Covid-19	-2	-1	1	1	0	-1	-1	1	2
	-3	-2	3	2	-1	-3	-2	-1	2
				-1	2	-1	0	0	0
			Courses	0	0	0	0	0	0
				1	-1	1	0	0	0
							0	0	0
					Student perf		0	0	0
							0	0	0

Expert 3 Target variables

	Courses			Student perf			Financial		
	2	1	-2	0	0	0	0	0	0
Covid-19	-3	2	3	-1	0	1	-1	0	1
	-3	2	3	-1	0	1	-2	-1	2
				0	0	0	0	0	0
			Courses	1	-1	1	0	0	0
				1	-1	1	0	0	0
							0	0	0
					Student perf		0	0	0
							0	0	0

Expert 4 Target variables

	Courses			Student perf			Financial		
	1	3	-3	0	0	0	0	0	0
Covid-19	-3	-2	3	1	-1	1	-1	0	1
	-3	-3	3	1	-1	1	-2	0	2
				0	0	0	0	0	0
			Courses	1	-1	1	0	0	0
				1	-1	1	0	0	0
							0	0	0
					Student perf		0	0	0
							0	0	0

Expert 5 Target variables

	Courses			Student perf			Financial		
	1	0	-1	-1	0	1	-2	2	1
Covid-19	-2	-1	2	1	-1	-2	-1	-1	2
	-3	1	3	2	1	-2	-2	-1	2
				0	2	-1	0	0	0
			Courses	1	0	1	0	0	0
				2	-1	2	0	0	0
							0	0	0
					Student perf		0	0	0
							0	0	0

Table A1: Utilities for the outcomes of the uncertainty factors.

Uncertainty factor	**Utility for outcome 1**	**Utility for outcome 2**	**Utility for outcome 3**
1	Stable: 0	Moderate: -0.4	Critical: -1
2	Normal: 0	Mixed: -0.1	Remote: -0.5
3	Low: -1	Normal: 0	High: 1
4	Impaired: -0.1	Normal: 0	Improved: 0.1

Table A2: Limits for the correction algorithm.

Pair	Lower bound	Upper bound
(COVID-19, Financial situation)	0.95	1.05
(Course arrangements,Student performance)	0.98	1.02