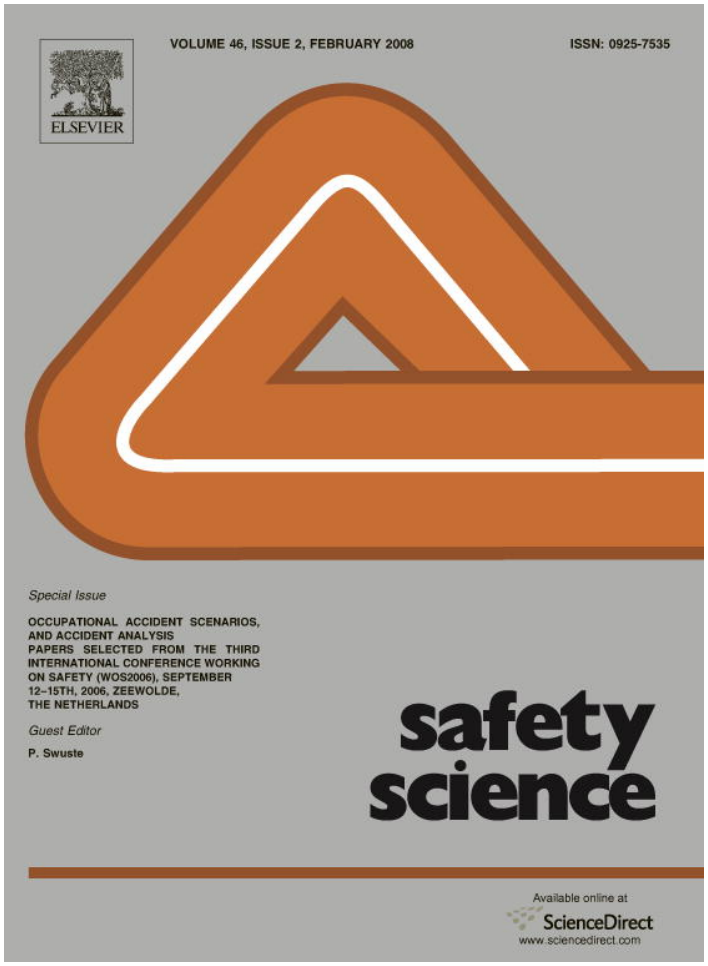

Available online at www.sciencedirect.com
ScienceDirect
Safety Science 46 (2008) 234–244
SAFETY SCIENCE
www.elsevier.com/locate/ssci

Fifteen years of expert judgement at TUDelft

L.H.J. Goossens [a,*], R.M. Cooke [b,c], A.R. Hale [a], Lj. Rodić-Wiersma [d]

[a] *Delft University of Technology, Delft, The Netherlands Safety Science Group, TBM/TUDelft, Jaffalaan 5, NL-2628 BX Delft, The Netherlands*
[b] *Delft University of Technology, Delft, The Netherlands, Department of Mathematics, Mekelweg 4 NL-2628 CD Delft, The Netherlands*
[c] *Resources for the Future, Washington, DC, USA*
[d] *UNESCO-IHE Institute for Water Education, Westvest 7, NL-2611 AX Delft, The Netherlands*

Received 3 April 2006; received in revised form 19 February 2007; accepted 30 March 2007

Abstract

Over the last fifteen Delft University of Technology (both the Safety Science Group and the Department of Mathematics of TUDelft) has developed methods and tools to support the formal application of expert judgement. Over 800 experts assessed over 4000 variables, in total representing more than 80,000 elicited questions. Applications were made in a variety of sectors, such as nuclear applications, the chemical and gas industries, toxicity of chemicals, external effects (pollution, waste disposal sites, inundation, volcano eruptions), aerospace sector and aviation sector, the occupational sector, the health sector, and the banking sector.

The techniques developed at TUDelft can be applied to give either quantitative assessments or just qualitative and comparative assessments. The application of these techniques is driven by a number of principles, including scrutability, fairness, neutrality, and performance control. The overall goal of these formal methods is to achieve rational consensus in the resulting assessments. Performance criteria are based on control assessments, that is, assessments of uncertain quantities, closely resembling the variables of interest, for which true values (e.g., from experiments) are known *post hoc*. The use of empirical control assessments is a distinctive feature of the Delft methods. A *Procedure Guide for Structured Expert Judgement* is published by the European Commission as EUR 18820. This paper highlights the comparative assessments for which the Safety Science Group was the prime responsible.

Keywords: Expert judgement

1. Introduction

Parameters necessary for modelling physical, chemical or biological behaviour are not always known with certainty. Experts may have valuable knowledge about models and parameters for problems in their specific field of interest. This knowledge is not certain, but is entertained with an implicit level of subjective confidence,

* Corresponding author.
E-mail addresses: l.h.j.goossens@tbm.tudelft.nl (L.H.J. Goossens), a.r.hale@tbm.tudelft.nl (A.R. Hale), l.rodic@unesco-ihe.org (Lj. Rodić-Wiersma).

0925-7535/$ - see front matter © 2007 Elsevier Ltd. All rights reserved.
doi:10.1016/j.ssci.2007.03.002

or *degree of belief*. The quantification and aggregation of experts' degrees of belief may provide important input to a decision maker, and may result in optimally defensible choices of parameters of models.

Behavioural and mathematical approaches are available for the elicitation and aggregation of individual experts' assessments (Cooke, 1991; Clemen and Winkler, 1999). Mathematical aggregation methods construct a single combined assessment per variable by applying procedures or analytical models that operate on the individual assessments. In contrast, behavioural aggregation methods involve interaction of the experts with a view to accomplishing homogeneity of information of relevance to the experts' assessments of the variables of interest. Through this interaction, some behavioural approaches, e.g., Kaplan's expert information approach (Kaplan, 1992), aim at obtaining agreement among the experts on the final probability density function obtained per variable. In others, e.g., approaches discussed by Budnitz et al. (1998) and by Keeney and Von Winterfeldt (1989) the interaction process is followed by simple mathematical combining, such as equal weighting, of the individual experts' assessments in order to obtain a single (aggregated) probability density function per variable. Fixed interaction procedures can be applied, or alternatively, the study team could design a dedicated procedure to suit a particular application. Both mathematical approaches with some modelling and behavioural approaches seem to provide results that are inferior to simple mathematical combination rules (Clemen and Winkler, 1999). Furthermore, a group of experts tends to perform better than the average solitary expert, but the best individual in the group often outperforms the group as a whole (Clemen and Winkler, 1999). This motivates the elicitation of the assessments of individual experts without any interaction, followed by mathematical aggregation in order to obtain a single assessment per variable, thereby weighting the individual experts' assessments based on their quality (Cooke, 1991).

2. Expert judgement and Safety Science Group

Over the last fifteen years Delft University of Technology has developed methods and tools to support the formal application of expert judgement. The development of one of the tools, EXCALIBUR software to aggregate experts' assessments, originated in the Safety Science Group at TUDelft and reached a mature state in the Mathematics Department of TUDelft (Cooke, 1991). The Ministry of Housing, Physical Planning and Environment granted a project in 1987, in which the opportunities of expert judgement were explored and practical models were made explicit (Goossens et al., 1989).

The techniques can be applied to give either quantitative assessments (the Classical Model) or only qualitative and comparative assessments (the paired comparisons model). The former give rise to assessments of uncertainty in the form of probability distributions, from which nominal values of parameters can be derived for practical applications; for the range of applications performed by TUDelft, see Table 1. The latter lead to rankings of alternatives; for the range of applications performed by TUDelft, see Table 2.

The application of these techniques is driven by a number of principles, including *scrutability* (all data and all processing tools are open to peer review and results must be reproducible by competent reviewers), *fairness* (experts are not pre-judged), *neutrality* (methods of elicitation and processing do not bias results), and *performance control* (quantitative assessments are subjected to empirical quality controls).

The overall goal of these formal methods is to achieve *rational consensus* in the resulting assessments. This requires that diverse stakeholders buy into the process by which the results are reached, and that the process itself optimises performance, as measured by valid performance criteria. Performance criteria are based on control assessments, that is, assessments of uncertain quantities, closely resembling the variables of interest, for which true values (e.g., from experiments) are known *post hoc*. Criteria for analysing control assessments are closely related to standard statistical methods, and are applied both to expert assessments, and to the combinations of expert assessments. The use of empirical control assessments is a distinctive feature of the Delft methods. A *Procedure Guide for Structured Expert Judgement* is published by the European Commission as EUR 18820 (Cooke and Goossens, 2000).

The resources required for an expert judgement study vary greatly depending on size and complexity of the study. A trained uncertainty analyst is required for defining the issues and processing the results. Past studies have used between four and twenty experts. The amount of expert time required for making the assessments depends on the subject and may vary between a few hours and a week, per expert. Total time required for studies in the past varies between one man-month to one man-year. Other variables determining the resource

Table 1
Overview of expert judgements performed with uncertainty estimates by the TUDelft team and liaisons

Sector	Number of experts	Number of variables	Number of elicitations
Nuclear applications	98	2203	20,461
Chemical and gas industries	32	217	3386
Chemical toxicity to humans	24	186	1105
Groundwater, and water pollution	18	59	497
Moveable barriers and Dike ring failures	31	153	3217
Volcano eruptions, and reliability of dams	231	673	29,079
Space shuttles, space debris, and aviation	51	161	1149
Health items: bovine respiratory diseases, Campylobacter on chickens, and SARS	46	240	2979
Banking issues: options, rents, and operational risks	24	119	4324
Occupational issues: falls from ladders, and thermal physics of buildings	13	70	800
Rest group	19	56	762
In total	587	4137	67,759

Table 2
Overview of expert judgements performed with paired comparisons by the TUDelft team

Sector	Number of experts	Number of variables	Number of elicitations
Chemical industries: safety management systems	114	83	5703
Chemical industries: flanges and valves	16	20	900
Reliability of landfill technologies	81	35	5871
Water pollution	82	64	2352
In total	293	202	14,826

commitment are travel, training given to experts in subjective probability assessments and level of documentation. Processing and write up of the results are greatly facilitated by software support.

This paper will highlight two cases for which the paired comparisons Model is used, as the applications with the Classical Model, for which the Safety Science Group was responsible, are already published in many publications (Cooke and Goossens, 2004; Goossens and Cooke, 1997, 2001; Goossens and Harper, 1998; Goossens and Kelly, 2000; Goossens et al., 1998; Van der Fels-Klerx et al., 2002, 2005).

3. Structured expert judgement

Expert judgement has always played a large role in science and engineering. Increasingly, expert judgement is recognised as just another type of scientific data, and methods are developed for treating it as such. In all cases, the judgements of more than one expert are elicited. The questions of measuring performance of experts and combining their judgements are addressed more fully in Cooke (1991).

In the world of engineering, technical expertise is generally separated from value judgements. Engineering judgement is often applied to bridge the gap between hard technical evidence and mathematical rules on the one hand and unknown characteristics of a technical system. Numerical data have to be derived suitable for the practical problem at hand. Engineers are quite able to provide these required engineering data which are essentially subjective data driven by engineering models and experience. The same is true for expert judgements. Engineering models and experience largely drive the subjective experts' assessments. That is why certain professionals become experts in certain fields of interest.

4. Paired comparisons

In the Paired Comparison method, experts are asked to rank alternatives pair wise according to some criterion like preference, beauty, feasibility, etc. If 20 items are involved in total, 190 comparisons must be made;

each item is compared with the 19 others. Since each item is compared with all the other items, there is a great deal of redundancy in the judgement data. Various processing methods are proposed for distilling a rank order from the pair wise comparison data. According to the method chosen and the availability of some measured values, the data can be further reduced to an interval or even a ratio scale. Paired comparisons were originally introduced for studying psychological responses (Thurstone, 1927), and have been applied to consumer research (Bradley, 1953), to the assessment of human error probabilities (Comer et al., 1984), to the assessment of failure probabilities (Goossens et al., 1989), to the assessment of landfill technology failures (Rodić, 2000), and to the assessment of safety management options (Hale et al., 1999, 2000). For a mathematical review, see David (1963). The method of paired comparisons yields no assessment of uncertainty. Methods for evaluating the degree of expert agreement and consistency are available.

5. Examples of expert judgements in the Safety Science Group

This section highlights some of the many expert judgement exercises executed with the Delft methods, where in particular the Safety Science Group was responsible for the overall project. The EU Seveso II Directive requires risk assessments for third party risks and for water pollution risks of chemical establishments. With respect to the water pollution risks, implementation of the Directive into the previous software tool (VERIS) required ranking of contributions of the management factors and relative failure frequencies of chemical activities (such as continuous and batch processes, tank storage, and loading and unloading of chemical substances) (Goossens and Cooke, 1997). Rankings of management factors were also assessed with the Paired Comparisons method for safety management systems (Hale et al., 1999, 2000) as well as for the reliability of engineered controls at solid waste landfills (Rodić, 2000; Rodić-Wiersma et al., 2001a,b). Both studies will be explored to a larger extent in this paper.

Other large expert judgement exercises carried out by the Safety Science Group aimed at providing quantitative assessments. Airborne releases of large amounts of toxic chemicals could result in huge fatal consequences. Dose-response relations of these chemicals are largely uncertain and difficult to achieve through animal experiments. Dose-response relations have been established for a few chemicals to show the potential of formal expert jugdement (Goossens et al., 1998). Third party risk assessments of chemical establishments in the Netherlands require dose-response relations represented by probit relations. In general, the choice of the probit parameters dominates the third party risk calculations heavily. For that reason, the choice of parameters has been taken as a joint responsibility between authorities and industries in the Netherlands.

In the succession of the last example quantitative assessments of probability distributions for important parameters in accident consequence models for nuclear power stations (Goossens and Harper, 1998; Goossens and Kelly, 2000) have been derived and used in the uncertainty analysis of the accident consequence software package COSYMA. In this particular project experts also assessed correlations between parameters as conditional probabilities. This large project jointly financed by the European Commission and the United States Nuclear Regulatory Commission resulted in the *Procedures Guide for Structured ExpertJudgment* (Cooke and Goossens, 2000), also known as EUR 18820. The *Procedures Guide* also describes the formal steps of expert judgement exercises in greater detail. A summary of the steps is provided in the next section.

In the veterinary sector expert judgements were used to come up with model parameters for bovine respiratory diseases of calves (Van der Fels-Klerx et al., 2002). The quantitative experts' assessments were fed into economic models for farming practices. Recently also the distribution of *Campylobacter* bacteria on chicken skin and meat were assessed with expert judgements (Van der Fels-Klerx et al., 2005).

6. Protocol and Procedures Guide

The Procedures Guide document (Cooke and Goossens, 2000) provides details of the protocol for a full expert judgement exercise. The protocol refers in particular to expert judgement exercises with the aim of achieving uncertainty distributions for uncertainty analyses. In that field of application the methods developed at Delft University of Technology have benefited from experiences gained with expert judgement in the United States with the NUREG-1150 protocol. For sake of clarity, the Procedures Guide represents a mix of these

developments and is not limited to NUREG-1150 type applications only. For paired comparisons exercises the same protocol is applied, whereby step (4) is omitted.

The protocol consists of 15 steps:

Preparation for elicitation:

(1) Definition of case structure
(2) Identification of target variables
(3) Identification of query variables
(4) Identification of performance variables
(5) Identification of experts
(6) Selection of experts
(7) Definition of elicitation format document
(8) Dry run exercise
(9) Expert training session

Elicitation

(10) Expert elicitation session

Post-elicitation

(11) Combination of expert assessments
(12) Discrepancy and robustness analysis
(13) Feed back
(14) Post-processing analyses
(15) Documentation

6.1. Case study 1: Reliability of landfill technologies

In this particular Ph.D.-project (Rodić, 2000) the main relevant questions were:

– How reliable is the technology currently applied at solid waste landfills?
– What are the most sensitive elements?
– What are the factors that determine the success of landfill technology performance?

For landfills, system failure can be defined in two distinct manners. One definition takes the breach of containment as the system failure. This means that any egress of substances constitutes failure of the waste containment system. The other definition states that system failure is an event of breach of containment such that subsequent release of contaminants could cause adverse health and/or environmental effects. In this study the former definition is adopted since the main goal of the study – generic investigation of performance of containment technology – requires identification of all possible causes of contaminant egress, regardless of the possible consequences.

6.2. Expert judgement

Rodić's literature study (Rodić, 2000) showed that the field data on landfill failures were insufficient for establishing probability distributions for failure events of individual physical components of a landfill containment system. Therefore, expert judgement data were necessary. During the preparatory phases of the study, it became apparent that experts were reluctant to give any quantitative estimates of the probability of the occurrence of individual failure events for various elements of landfill technology. They said that there had not been sufficient field evidence on which to base such estimates. Therefore, experts were subsequently asked to give

their opinions not about the absolute probability of occurrence of the failure events, but the likelihood of occurrence of these events relative to one another.

6.3. Selection of experts

There are no objective criteria for proclaiming a person an expert. Experts ought to be selected as objectively as possible, i.e. in cases where the number of potential experts is large world-wide (over 300 experts), the opinion should be asked of those professionals who really stand out for their knowledge and experience. So the first selection of experts by peer designations is considered the best available. Accordingly, in this study experts were identified by their colleagues. For that purpose, nominations were solicited from professionals in the field of landfill technology and management. The procedure followed was similar to that used in the study by Hawkins and Graham (1988). The names of the professionals were collected from the literature consulted in the previous phases of the study. A letter was sent to each person (author of an article) with the request to nominate up to six individuals with expertise on landfill technology. Of the approximately 340 specialists in Europe and North America to whom the letters were sent, 120 wrote back with their recommendations.

The experts with most nomination counts were then contacted and asked to participate in the main part of the study. At equal counts, care was taken that experts from different types of institutions were included. Virtually all experts who were contacted at this stage showed remarkable readiness to cooperate and share their views and opinions. None of the experts asked for financial reciprocation for his participation in this study. The participating experts originated from Austria, Belgium, Canada, France, Germany, Italy, the Netherlands, Switzerland, the United Kingdom and the United States. Furthermore, they came from the following types of institutions: research institutes; ministries of the environment and governmental environmental agencies; engineering consultancies providing landfill design; consultancies providing construction quality control and assurance; manufacturing industries of geosynthetic materials; and waste treatment and waste disposal industries.

6.4. Selection of questions

Questions to be submitted to the experts were defined only after extensive preparations consisting of the following activities:

- literature search and study,
- preliminary interviews about the list of failure events,
- discussion of general landfill issues,
- preparation of paired comparisons questionnaires.

In order to establish the state of (published) knowledge about landfill technology performance, a literature study was carried out (Rodić, 2000). On the basis of the literature studied, a detailed inventory of failure events was made. Subsequently, preliminary interviews with specialists, mostly from the Netherlands, were conducted to verify that the list of basic events encompassed all possibly relevant points. Besides contacting researchers, designers and policy specialists, an effort was also made to consult with people in landfill operations whose experience is often unpublished. Furthermore, not only specialists in landfill technology were contacted but also those specialised in complementary scientific and technological branches (such as geology, geochemistry, soil chemistry and waste treatment).

In the course of the discussion of the assembled list of landfill failure events with the specialists in the preliminary interviews, some questions emerged that could not be directly linked to the failure events. They concerned more fundamental issues of waste landfilling rather than detailed features of the landfill technology currently applied. It was necessary to address these issues before entering the following phase of the project – formal elicitation of expert opinions. Examination of these fundamental landfill issues turned out to be a very fruitful exercise. It discerned the most problematic areas and thereby enabled clustering the failure events from the list in a conveniently compact and yet comprehensive manner.

On the basis of the assembled list of basic failure events and on the basis of experts' answers to the fundamental issues, the items (events) were defined for the following stage of the study – the formal elicitation of expert opinions by paired comparisons.

6.5. Formal elicitation of expert judgements by Paired Comparisons

Using the method of paired comparisons, failure events are put in pairs and the experts are asked to point to the one which, in their opinion, would be more likely to occur. As mentioned earlier, the strength of this method lies in its redundancy, because each item is evaluated several times, precisely $n - 1$ times for n items. (Reverse order pairs are not included.) Thereby a good picture is formed of experts' internal consistency. At the same time, the overall judgement is not sensitive to an occasional error in reasoning on the part of the experts. In total, 36 experts participated in this main stage of the study. Some gave their opinion about the performance of landfill liners only, some about the leachate collection and removal system, and some about both, according to their field of expertise.

In this study experts were explicitly asked to give their opinion based simply on the real situation as it is in practise, and not confine themselves to the best design companies, the best testing institutes, the best contractors, and the best site operators. Furthermore, they were asked to give their opinion about the likelihood of occurrence of the items compared regardless of the extent of the possible consequences of the event (i.e. not to choose the item which, should it occur, would have more dramatic consequences, but the one which would be more likely to occur). Also experts were regularly asked for their arguments for the preferences they had given.

6.6. Expert judgement results

As an example, the results of the assessments for failures of landfill liners in the short term are explained. The questionnaire was filled in by 36 experts. Four experts exceeded the consistency criteria and their assessments were removed from the expert judgement database. As shown in Table 3 the experts found that most problems stem from construction (installation); the two highest scoring items belong to the construction phase, namely:

- bad quality of the geomembrane seams and/or bad quality of the clay compaction,
- ordinary mechanical damage during installation.

Due to its pronounced chemical resistance, high-density polyethylene (HDPE) has become the material of choice for geomembrane liners in landfill applications. However, HDPE sheets are very inflexible and therefore very difficult to work with. Seaming (welding) of the geomembrane sheets in the field is far from easy. Geomembrane seaming is particularly difficult around penetrating objects such as pipes. Among numerous factors that influence the seam quality, several are readily controllable: temperature, rate of movement and pressure. Although sophisticated monitoring equipment has been added to the welding devices, it has turned out to be difficult to maintain the values of these parameters within the desired range throughout the seaming process in practice. (In addition to these, cleanliness of the bonding surfaces – absence of dirt and moisture – is very important and is fairly straightforward to control.) As a part of the solution, in addition to advanced equipment, higher training requirements and certification programmes for welding technicians have been introduced. Also, reward/punishment motivation strategies have been devised in an attempt to achieve and maintain the quality of welding, but without consistent success. Bearing all this in mind, it seems unlikely that the reason for the problem lies primarily in the lack of scientific insights or deficiencies of the equipment used. The reason should probably be sought in difficult working conditions under which welding is performed, such as long working days, work on slopes, hot weather, monotonous work, constant division of attention between the geomembrane seams being welded and the monitors, etc.

Another problem with geomembrane seam quality is that "at present there are no universally accepted criteria that define a 'good' seam" (Peggs, 1994). It is not clear how the currently used criteria relate to the actual (long-term) performance of the seams. The currently performed tests cannot discern the presence of microscopic defects introduced into the geomembrane's molecular structure by application of heat and pressure

Table 3
Expert judgement results for failures of landfill bottom liners in the short term (the score represents the fraction of contribution to the overall failure frequency of the bottom liners)

Item	Brief description	Score
4	Bad geomembrane seams and/or clay compaction	0.24
3	Installation damage	0.20
10	Not safeguarding liner in operation	0.14
12	Pipes penetrating liner	0.08
2	Bad quality material accepted at site	0.07
7	Failure of LCRS	0.07
1	Bad design and/or choice of materials	0.07
9	Financial shortcuts in installation	0.03
5	Geotechnical failure	0.03
8	Breach by vertical pipes	0.03
11	Inadequate siting	0.02
6	Unanticipated chemical attack	0.02

during the welding process. These defects may, however, adversely affect the long-term performance of the seam (Rollin et al., 1990).

6.7. Case study 2: Safety management systems and expert judgements

This part of the paper describes the results of a study aimed at deriving measures of the relative importance of management factors on risk control in the chemical industries, in particular related to maintenance management in major hazard chemical plants. Two projects (Hale et al., 1999, 2000) were conducted within the framework of the I-Risk project (Oh et al., 1998). They used the management model developed in that project as the basis for deriving a protocol. Experts judged the relative importance of eight generic management areas which determine the quality of completion of safety critical tasks:

– availability of suitable personnel,
– competence of those personnel,
– their commitment to safety,
– communication and coordination,
– conflict resolution (priority of safety vs. other goals),
– interface design,
– procedures and plans,
– delivery of correct spares and replacements.

6.8. Expert judgement

The judgements were made in respect of the management of eight parameters found in the I-Risk model (Papazoglou and Aneziris, 1998):

– time for maintenance – Tm,
– time for repair – Tr,
– maintenance interval – Im,
– test interval – It,
– probability of failure to replace like with like in maintenance – L/L,
– respect of equipment design envelope during maintenance – RDe,
– human error in maintenance – HEm,
– human error in inspection – HEi.

In the course of the study the parameters Im and It were combined on the advice of the experts, as they were felt to share management influences. This left seven parameters, all related to inspection and maintenance management activities, to be judged on the eight influences. In addition the experts were asked to judge the relative contribution of failure reasons from eight domains of management to the overall hardware failure rate. Three of those domains related directly to parameters above (Im, L/L, RDe), allowing us to link these parameters through to their effect on failure rates.

6.9. Selection of experts

The experts used for making the judgements were recruited from 14 Dutch major hazard plants, through the network of contacts of the project leaders. All were, or had been, managers of inspection and/or maintenance functions in major hazard plants for at least 3 years (average 15 years). In the first study (first four parameters from the list above) 18 experts took part, in the second study (the remaining five parameters) 12 experts. Five experts took part in both studies.

6.10. Selection of questions

The method used is described in detail in Costa (1998) and summarised in Hale et al. (1999). A list of potential management factors of influence on each parameter was produced from the literature and discussion with experts, within the eight generic categories of influence listed above. After refinement of the lists to reduce them to manageable proportions, between 7 and 13 influences remained over per parameter. Not all types of influence were included in each parameter, since the experts consulted found some irrelevant (see Table 4). Definitions of each parameter and the scenarios leading to poor management of each were also prepared. The definitions, scenarios and list of influences were sent to the experts in advance to help them orient themselves to their task. They were discussed at the start of the elicitation sessions to clarify any issues about what was being asked. The experts were given the chance to add influences not on the list, which occurred for four of the parameters. The influences were then presented in a questionnaire, pairing each influence with each other and asking the experts to respond simply by circling the more important in each pair. Nine elicitation sessions were held with between one and seven experts present.

6.11. Formal elicitation of expert judgements by paired comparisons

The data was analysed using the COMPAIR program (Cooke and Solomatine, 1992) and tested for inconsistencies by analysing the number of circular triads ($A > B > C > A$). Experts with more than a threshold number of circular triads on a given parameter (dependent on the number of paired comparisons to be made on the parameter) were removed from the analysis. This resulted in the loss of up to four experts on a given parameter.

Table 4
Percentage influence and coefficients. In brackets number of influences

Parameter influence	Tm	Tr	Im/It	L/L	RDe	HEm	HEi
Availability	0 (0)	7.6 (1)	0 (0)	3.8 (1)	2.5 (1)	3.0 (1)	4.0 (1)
Competence	40.3 (4)	29.3 (4)	37.5 (3)	40.5 (2)	40.7 (1)	24.9 (3)	39.3 (2)
Commitment	6.5 (1)	5.9 (1)	16.3 (2)	10.3 (1)	8.1 (1)	8.2 (2)	8.8 (1)
Communication	19.7 (2)	26.7 (2)	2.7 (1)	0 (0)	24.6 (1)	33.4 (1)	24.8 (1)
Conflict resolution	0 (1)	0 (1)	1.6 (1)	0 (0)	8.1 (1)	10.2 (2)	6.5 (1)
Interface	0 (0)	0 (1)	3.0 (1)	3.3 (1)	5.8 (1)	2.6 (1)	7.9 (2)
Procedures/plans	33.5 (3)	13.7 (3)	38.9 (4)	26.1 (2)	10.6 (1)	14.7 (1)	9.0 (1)
Spares	0 (1)	16.8 (1)	0 (0)	16.0 (1)	0 (0)	0 (0)	0 (0)
No. of experts	14 of 18	16 of 18	14 of 18	11 of 12	10 of 12	12 of 12	8 of 12
Coëfficiënt u	0.065	0.203	0.259	0.137	0.211	0.203	0.165
Coefficient W	0.132	0.318	0.393	0.261	0.348	0.408	0.269

Table 5
Distribution of hardware failure influences

Area of management control	Percentage influence
Change (configuration) management	47.0
Operations	16.9
Engineering design	11.5
Process	10.8
Maintenance execution (RDe)	8.6
Maintenance interval (Im)	3.6
Spares (L/L)	1.0
Raw materials	0.6
No. of experts	11
Coefficient u	0.417
Coefficient W	0.564

6.12. Expert judgements results

The program assigns rank orders and weights to the influences depending on how often they are rated as the most important in a pair. Coefficients of agreement and concordance were calculated to indicate the agreement between the rank orders and between the ratings of each pair of items across all the experts. Values of at least 0.2 (on a scale from 0 to 1) are considered necessary before it can be concluded that there is a reasonable consensus across an expert group (Cooke, 1991).

Table 4 gives the results of the weighted importance (%) of the eight management areas for the parameters studied, and the coefficients of agreement and concordance found across the whole group of N experts (those with too many circular triads having been removed). Table 5 shows the ratings for the areas of management implicated in the hardware failures.

The coefficients are acceptable (though sometimes only just) on both coefficients for 5 of the 8 parameters judged and on one of the two coefficients for two others. The low agreement on time for maintenance seems explicable from the reports of the participants that they found the judgement to be made was unrealistic because safety critical plant, even if redundant, was not maintained while the plant was on line (Hale et al., 1999). The poor performance on "replacing like with like" may be caused by the lack of agreement from the experts with only inspection management experience, who may have insufficient direct experience to judge the influences. If we take only the experts with maintenance management experience, the coefficients rise to 0.252 and 0.440, respectively. Unfortunately, for the parameter "human error in inspection", where one would expect the managers with inspection management experience to produce a better coefficient of agreement than those with only maintenance management experience, they do not. Again the maintenance managers are the more consistent among themselves. However numbers are too small to place much faith in sub-analyses.

7. Conclusions

Valid methods for eliciting expert judgements have been developed at Delft University of Technology. In the past fifteen years over 80,000 elicitations were made by almost 900 experts. The methods have proven to be mature and provide a scientific tool for achieving additional data that would otherwise remain unavailable. Two examples of case studies, where the project outcome was driven by expert judgements, are described to illustrate the extent to which the methods may lead. The case studies apply the paired comparisons method and show how expert judgements provide additional information to come to useful answers in projects where the parameters are uncertain.

References

Bradley, R., 1953. Some statistical methods in taste testing and quality evaluation. Biometrica 9, 22–38.

Budnitz, R.J., Apostolakis, G., Boore, D.M., Cluff, L.S., Coppersmith, K.J., Cornell, C.A., Morris, P.A., 1998. Use of technical expert panels: applications to probabilistic seismic hazard analysis. Risk Analysis 18, 463–469.

Clemen, R.T., Winkler, R.L., 1999. Combining probability distributions from experts in risk analysis. Risk Analysis 19, 187–203.
Comer, K., Seaver, D., Stillwell, W., Gaddy, C., 1984. Generating human reliability estimates using expert judgement. NUREG/CR-3688.
Cooke, R.M., 1991. Experts in uncertainty. Opinion and subjective probability in science. Oxford University Press, New York/Oxford.
Cooke, R.M., Goossens, L.H.J., 2000. Procedures guide for structured expert judgement. European Commission. Report EUR 18820.
Cooke, R.M., Goossens, L.H.J., 2004. Expert judgement elicitation for risk assessments of critical infrastructures. Journal of Risk Research 7, 643–656.
Cooke, R.M., Solomatine, D., 1992. EXCALIBR – integrated system for processing expert judgements, Version 3.0, User's manual. Delft University of Technology and SoLogic, Delft.
Costa, M.A.F., 1998. Relative weight of maintenance management influences on technical risk parameters. Graduation report. Safety Science Group, Delft University of Technology, Delft and University of Minho, Minho.
David, H., 1963. The Method of Paired Comparisons. Charles Griffin.
Goossens, L.H.J., Cooke, R.M., 1997. Applications of some risk assessment techniques: formal expert judgement and accident sequence precursors. Safety Science 26, 35–47.
Goossens, L.H.J., Cooke, R.M., 2001. Expert judgement elicitation in risk assessment. In: Linkov, I., Palma-Oliveira, J. (Eds.), Assessment and management of environmental risks. Kluwer Academic Publishers, Netherlands, pp. 411–426.
Goossens, L.H.J., Harper, F.T., 1998. Joint EC/USNRC expert judgement driven radiological protection uncertainty analysis. Journal of Radiological Protection 18, 249–264.
Goossens, L.H.J., Kelly, G.N., 2000. Expert judgement and accident consequence uncertainty analysis. Special Issue. Radiation Protection Dosimetry 90, pp. 293–381.
Goossens, L.H.J., Cooke, R.M., van Steen, J., 1989. Expert opinions in safety studies. In: Philosophy and Technical Social Sciences, vols. 1–5. Delft University of Technology, Delft.
Goossens, L.H.J., Cooke, R.M., Woudenberg, F., van der Torn, P., 1998. Expert judgement and lethal toxicity of inhaled chemicals. Journal of Risk Research 1, 117–133.
Hale, A.R., Costa, M.A.F., Goossens, L.H.J., Smit, K., 1999. Relative importance of maintenance management influences on equipment failure and availability in relation to major hazards. In: Schuëller, G.I., Kafka, P. (Eds.), Safety and Reliability, ESREL '99, A.A. Balkema, Rotterdam, pp. 1327–1332.
Hale, A.R., Goossens, L.H.J., Costa, M.F., Matos, L., Wielaard, P., Smit, K., 2000. Expert judgement in the assessment of the contribution of safety management aspects to the control of major hazard risk. In: Kondo, S., Furuta, K. (Eds.), PSAM5 – Probabilistic Safety Assessment and Management, (PSAM5, Osaka, Japan, 27 November–1 December 2000). Universal Academy Press, Inc, Tokyo, Japan, pp. 1139–1144.
Hawkins, N.C., Graham, J.D., 1988. Expert scientific judgement and cancer risk assessment: a pilot study of pharmacokinetic data. Risk Analysis 8, 615.
Kaplan, S., 1992. Expert information' versus 'expert opinions'. Another approach to the problem of eliciting/combining/using expert knowledge in PRA. Reliability Engineering and System Safety 35, 61–72.
Keeney, R.L., Von Winterfeldt, D., 1989. On the uses of expert judgment on complex technical problems. IEEE Transactions on Engineering Management 36, 83–86.
Oh, J.I.H., Brouwer, W.G.J., Bellamy, L.J., Hale, A.R., Ale, B.J.M., Papazoglou, I.A., 1998. The I-Risk project: development of an integrated technical and management risk control and monitoring methodology for managing and quantifying on-site and off-site risks. In: Mosleh, A., Bari, R.A. (Eds.), Probabilistic Safety Assessment and Management. Springer, London, pp. 2485–2491.
Papazoglou, I.A., Aneziris, O.N., 1998. System performance modeling for quantification of organisational factors in chemical installations. In: Mosleh, A., Bari, R.A. (Eds.), Probabilistic Safety Assessment and Management. Springer, London, pp. 2093–2098.
Peggs, I.D., 1994. HDPE geomembrane seams: acceptance criteria and critical defects. In: Koemer, R.M., Wilson-Fahmy, R.F. (Eds.), Geosynthetic Liner Systems: Innovations, Concerns and Designs. Industrial Fabrics Association International, St. Paul, MN.
Rodić, Lj., 2000. Reliability of landfill technology. Ph.D. thesis. Delft University of Technology, Delft.
Rodić-Wiersma, Lj, Goossens, L.H.J., 2001. Assessment of landfill technology failure. In: Christensen, T.H., Cossu, R., Stegmann, R. (Eds.), Sardinia 2001a, 8th International Waste Management and Landfill Symposium, 1–5 October 2001, Sardinia, vol. I, pp. 605–704.
Rodić-Wiersma, Lj, Goossens, L.H.J., 2001. Landfill barrier technology performance: more than technology alone. In: Christensen, T.H., Cossu, R., Stegmann, R. (Eds.), Sardinia 2001, 8th International Waste Management and Landfill Symposium, 1–5 October 2001b, Sardinia, vol. III, pp. 93–102.
Rollin, A.L., Vidovic, A., Denis, R., Marcotte, M., 1990. Microscopic evaluation of high-density polyethylene geomembrane field-welding techniques. In: Peggs, I.D. (Ed.), Geosynthetics: Microstructure and Performance, ASTM STP 1076. American Society for Testing and Materials, Philadelphia, PA.
Thurstone, L.L., 1927. A law of comparative judgment. Psychological Review 34, 273–286.
Van der Fels-Klerx, H.J., Cooke, R.M., Nauta, M.N., Goossens, L.H.J., Havelaar, A., 2005. A structured expert judgment study for a model of Campylobacter transmission during broiler-chicken processing. Risk Analysis 25, 109–124.
Van der Fels-Klerx, H.J., Goossens, L.H.J., Saatkamp, H.W., Horst, S.H.S., 2002. Elicitation of quantitative data from a heterogeneous expert panel: formal process and application in animal health. Risk Analysis 22, 67–81.