Available online at www.sciencedirect.com

SCIENCE DIRECT®

Decision Support Systems 41 (2005) 84–111

Decision Support Systems

www.elsevier.com/locate/dsw

Design and evaluation of an intelligent decision support system for nuclear emergencies

K.N. Papamichail*, S. French

Information Systems Group, Manchester Business School, Booth Street West, Manchester M15 6PB, UK

Received 1 September 2002; accepted 1 April 2004
Available online 14 July 2004

Abstract

Intelligent Decision Support Systems (DSSs) use expert systems technology to enhance the capabilities of decision makers (DMs) in understanding a decision problem and selecting a sound alternative. Because of the people-centred focus of such technologies, it is important not only to assess their technical aspects and overall performance but also to seek the views of potential users. This paper draws from the literature to classify methods for assessing intelligent Decision Support Systems and discusses our experiences in developing, operating and evaluating an intelligent decision support system for nuclear emergencies. The system assists decision makers in the formulation and ranking of alternatives and communicates its recommendation in a natural language form. The application highlights insights from the development process and shortcomings of existing assessment methods. Lessons learned from the study, challenges encountered and recommendations for future practices are discussed.

Keywords: Assessment; Emergency management; ESY; Expert systems; Evaluation; Intelligent Decision Support Systems; Radiation accident; RODOS; Verification and validation methods

1. Introduction

Evaluation is an important activity that is often omitted during the development of a decision support system (DSS) or expert system. Even when evaluation is conducted, it is not undertaken throughout the development cycle of a system but rather at the end [1]. However, if deficiencies are identified early on in the system's development cycle then corrective actions can be taken that are easier and less expensive to carry out in the early phases rather than in later phases [7]. A thorough examination of a DSS would allow the system developers to find out how well the system works, how sound its advice is, and whether it addresses the needs of its users.

Evaluation is usually conducted in order to verify and validate a DSS. The terms 'verification' and 'validation' often have overlapping and interchangeable meanings in the literature [59]. According to

* Corresponding author. Tel.: +44 161 275 6539; fax: +44 161 275 6489.
E-mail addresses: nadia.papamichail@mbs.ac.uk (K.N. Papamichail), simon.french@mbs.ac.uk (S. French).

0167-9236/$ - see front matter © 2004 Elsevier B.V. All rights reserved.
doi:10.1016/j.dss.2004.04.014

Miser and Quade [59], "Verification is the process by which the analyst assures himself and others that the actual model that has been constructed is indeed the one he intended to build" whereas "Validation is the process by which the analyst assures himself and others that a model is a representation of the phenomena being modelled and that is adequate for the purposes of the study of which it is a part". O'Keefe et al. [67] give a rather shorter definition: "Verification is building the system right, validation is building the right system". "Verification is part of validation; a system that has not been built right is unlikely to be the right system" [66].

This paper describes the evaluation of the Evaluation subSYstem (ESY)—an intelligent DSS for nuclear emergencies [70,72]. The ESY is a module of Real-time Online DecisiOn Support system (RODOS) that provides comprehensive decision support in radiation accidents [4,23]. While other RODOS modules predict the radiological situation and calculate the consequences (e.g. health effects, cost) of countermeasures such as evacuation, sheltering and agricultural measures, the ESY compares strategies, i.e. combinations of countermeasures applied to areas affected by radiation. The ESY consists of three components:

- A Coarse Expert System—CES—that generates feasible strategies that satisfy several constraints [71].
- A Ranking Module that ranks alternative strategies based on their consequences and the preferences of the decision makers (DMs) [72].
- A Fine Expert System—FES—that outputs natural language reports to explain the ranking of the strategies, interpret sensitivity analysis graphs and identify the most important factors in the choice between two alternatives [73].

The ESY is an intelligent assistant that undertakes tasks such as generation and evaluation of alternatives and communicates its conclusions in a natural language form. Intelligent Decision Support Systems (DSSs) are interactive computer-based systems that use data, expert knowledge and models for supporting DMs in organisations to solve semi-structured problems by incorporating artificial intelligence techniques [84]. They draw on ideas from diverse disciplines such as decision analysis, artificial intelligence, knowledge-based systems and systems engineering. A review of intelligent DSSs that combine mathematical modelling with knowledge-based systems can be found in Silverman [89,90].

In order to assess the ESY, we have drawn from the literature and identified best practice in evaluating DSSs and expert systems. We have devised a strategy for assessing intelligent DSSs such as the ESY that involves the following assessment levels:

- *Technical verification*, i.e. looking inside the 'black box' to eliminate coding errors and check how well the system has been built, how accurate its output is and whether its advice is sound.
- *Performance validation*, i.e. assessing performance aspects of the system such as how well it works and performs its tasks and how accurate and complete its knowledge base is.
- *Subjective assessment*, i.e. collecting opinions to measure the utility of the system, establish whether it addresses the needs of its users and assess how well its interface is designed.

Preliminary results of the ESY assessment are reported in Papamichail [69]. Quality assurance guidelines for all the RODOS modules are given in Ranyard [79].

The ESY combines DSSs and expert systems technologies (in this paper the 'expert systems' and 'knowledge-based systems' terms are used interchangeably even though some differences between the terms have been identified elsewhere, e.g. Turban and Aronson [95]). However, a review of the literature shows that research on the verification and validation of decision-aiding tools often focuses on either the evaluation of DSSs or the appraisal of expert systems. This paper attempts to combine diverse methods into a unified assessment framework and discusses several challenges encountered during the ESY evaluation.

The objectives of this paper are:

- to discuss the design and development of an intelligent DSS for nuclear emergencies;
- to review the literature on the evaluation of DSSs and expert systems;
- to describe the assessment of the ESY; and,

- to draw conclusions and highlight the difficulties encountered during the development and evaluation of intelligent DSSs.

The paper is structured as follows. Section 2 describes the evaluation system—ESY and its components. A review of the literature on the evaluation of DSSs and knowledge-based systems is given in Section 3. The assessment of the ESY is described in two parts: technical verification and performance validation (Section 4) and subjective assessment (Section 5). The challenges encountered during the development and evaluation of the ESY are discussed in Section 6. Section 7 gives the conclusions of the study.

2. ESY—an evaluation system for nuclear emergencies

2.1. The decision-analysis process

Decision analysis can be seen as a consultation process that attempts to focus the attention of a decision maker (DM) on the important aspects of a decision problem. As with any other consultation, 'it starts with the definition of a decision problem and it ends with a commitment to an action plan' [80]. The decision process can be decomposed into three stages [39], as shown in Fig. 1 (a decision model is defined as a 'formal representation of the decision problem that reflects the DM's real situation'):

- Formulation of the decision model that reflects the decision problem, i.e. generating alternatives and identifying evaluation criteria.
- Evaluation of the decision model, i.e. computing the implications of the decision model, evaluating it using a formal decision method and producing a recommendation.
- Appraisal of the recommendation, i.e. analysing the recommendation and presenting the interpretation in a natural language form.

Feedback or refinement paths are provided to allow DMs to reevaluate the decision model or modify its formulation. The decision model is progressively refined until a DM is confident that the components, structure and values of the decision model accurately represent the decision problem [58]. Philips [77] argues that these final decision models are requisite because they are detailed enough to help DMs make a decision without considerable effort.

The ESY has been designed to support the decision analysis process (as presented in Fig. 1) in the event of a radiation accident. Each stage of the process is supported by an ESY component. The system is intended to support users ranging from scientists to emergency planning officers, health officials and politicians. At the beginning of a decision analysis, DMs might be confused and not confident about their preferences. The initial decision model might be a poor reflection of their values. As the analysis proceeds, DMs can interact with the ESY and receive feedback in the form of sensitivity analysis results and explanations. The ESY output (a ranked list of alternatives and explanations) helps them gain insight into the decision problem and in doing so revise their decision model. The interface allows DMs to refine their decision model by considering new alternatives (Coarse Expert System), changing their preference values and evaluation criteria (Ranking Module) and repeating the ranking of alternatives (Ranking Mod-

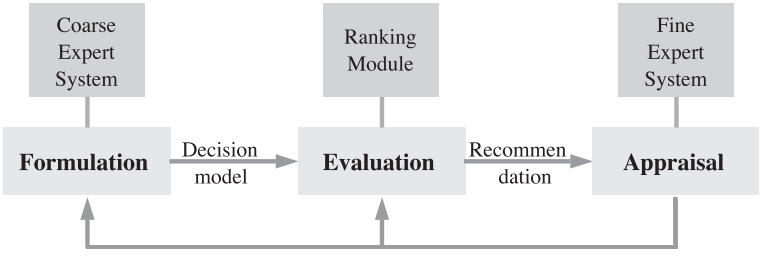

Fig. 1. The stages of a decision analysis process.

ule). This process is repeated until the DMs are satisfied with the output of the ESY. They are then ready to take a decision based on their own requisite decision model.

Fig. 2 illustrates the architecture of the ESY and the interactions expected and supported with DMs and experts. Unlike other radiation protection decision-aiding tools, the role of the ESY is not limited to processing and presenting information. It interactively guides DMs through the formulation, evaluation and appraisal stages of a decision analysis. The system devises and evaluates a decision model that is tailored to the preferences of the DMs. Its advice is prescriptive in the sense that it recommends the most preferred strategies based on the input of the DMs.

The protective measures and criteria taken into account in the evaluation of strategies change over time depending on the phase of the nuclear emergency. The ESY framework adjusts to the requirements of each phase. In the early phases of a nuclear accident (hours and days after the radiation accident), DMs are concerned with early-phase countermeasures (e.g. evacuation and sheltering) and criteria such as health effects, feasibility matters and to a lesser extent cost-related issues. In later phases (months and years after the accident), DMs have more time to spend on formulating strategies such as combinations of agricultural countermeasures and balancing both the short- and long-term health effects with the cost of the strategies.

2.2. Coarse expert system

In the early phases of a nuclear emergency, decisions have to be taken under time pressure and stress. DMs consider countermeasures such as evacuation and sheltering and devise a strategy, i.e. a portfolio of countermeasures to mitigate the consequences of the radiation accident. However, if they informally and intuitively try to identify strategies, they may come up with poor options. Support is therefore needed to ensure that superior and feasible alternatives are considered for evaluation.

We have developed a Coarse Expert System—CES—to automate the process of generating and identifying superior alternatives, which is often

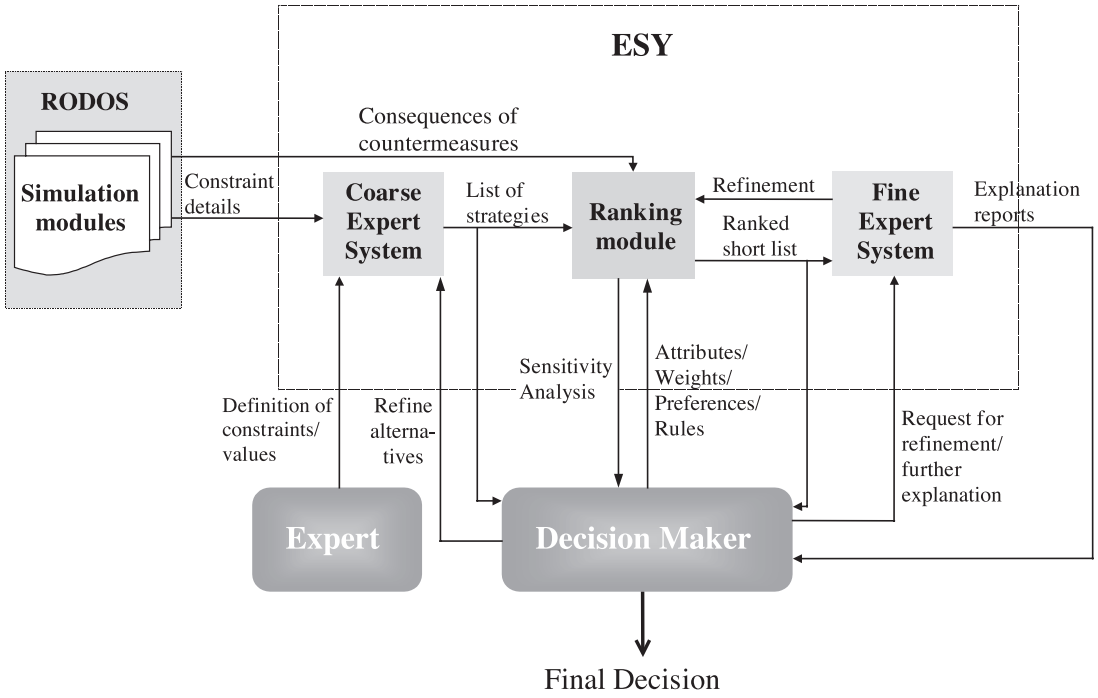

Fig. 2. The ESY architecture.

called screening [46]. In a nuclear emergency, the computational complexity of the screening problem grows exponentially with the number of contaminated areas and countermeasures available. The Coarse Expert System applies constraints defined by experts and recommendations provided by international radiation protection bodies. The criteria taken into account to screen alternatives include health safety, social and feasibility factors. For example, intervention is strongly recommended when the radioactive dose received by the population exceeds a given level. Temporal constraints do not allow issuing iodine tablets to people who have already been evacuated. Other rules ensure that neighbouring populations are treated in a continuous or similar way.

Fig. 2 illustrates the interactions expected and supported with experts and DMs. Experts can be consulted prior to the installation of the Coarse Expert System to choose which constraints and measures to consider as well as determine default values for the constraints. In the event of a radiation accident, scientists in discussion with DMs can specify the areas affected by radiation, define intervention levels and input the start and end times of implementing measurements.

Once the constraints are defined, the system runs to identify feasible strategies. The problem of generating and identifying feasible alternatives is represented as a constraint satisfaction problem. We have found that it was conceptually easy to model the search space of the screening problem using constraint programming [93]. The main attractiveness of this method over other approaches is that it represents a problem separately from the algorithms used to solve it and uses search algorithms that can take advantage of the particular features and intricacies of the problem at hand. The system provides real-time advice and considerably decreases the number of alternatives under consideration to a manageable fraction. Its output is a list of feasible strategies that are worthy of further evaluation.

DMs can view the strategies generated by the Coarse Expert System through a graphical user interface (see Fig. 3). The area around the source of radiation is typically divided into emergency planning blocks. Each strategy involves the implementation of a measure or a combination of measures in the affected blocks. The system illustrates each alternative in two forms: a graphic display and a text description. Different colours are used to depict a combination of measures in the affected areas.

2.3. Ranking module

After exploring strategies in the Coarse Expert System, a list of feasible alternatives is passed into the Ranking Module for further evaluation. The strategies are ranked based on their consequences and the preferences of the DMs. We have organised elicitation exercises to examine how DMs take decisions in radiation accidents and concluded that they find it very difficult to articulate which factors drive their decision making [4]. Their main objective is the return to normal living conditions. Other sub-criteria they may consider during the pre-release phase (i.e. when there is a suspicion for a release of radioactivity) as well as at the early phases of a nuclear emergency (i.e. hours or days after the radiation accident) are the following:

- *Collective dose*. This is the sum of the individual doses over a population. If large populations are exposed to radiation, then stochastic effects can be expected to occur even if the radiation levels and therefore individual doses are low. DMs were particularly interested in this attribute during the elicitation exercises.
- *Individual dose*. This is the sum of the equivalent doses individuals receive in all their tissues and organs. Although collective dose seems to be the primary concern of the DMs, this is still an important attribute because some individuals may be at a higher risk of being affected by deterministic or stochastic effects than the rest of the population.
- *Population number*. The number of people potentially affected by the radiation accident and involved in a strategy. This is a proxy attribute used because of difficulties in measuring the public acceptability of a strategy, the stress caused to the affected population and the feasibility of the measures taken at the early phases of a nuclear emergency.
- *Cost*. Few DMs have considered it at all in exercises and none did so significantly. Nonethe-

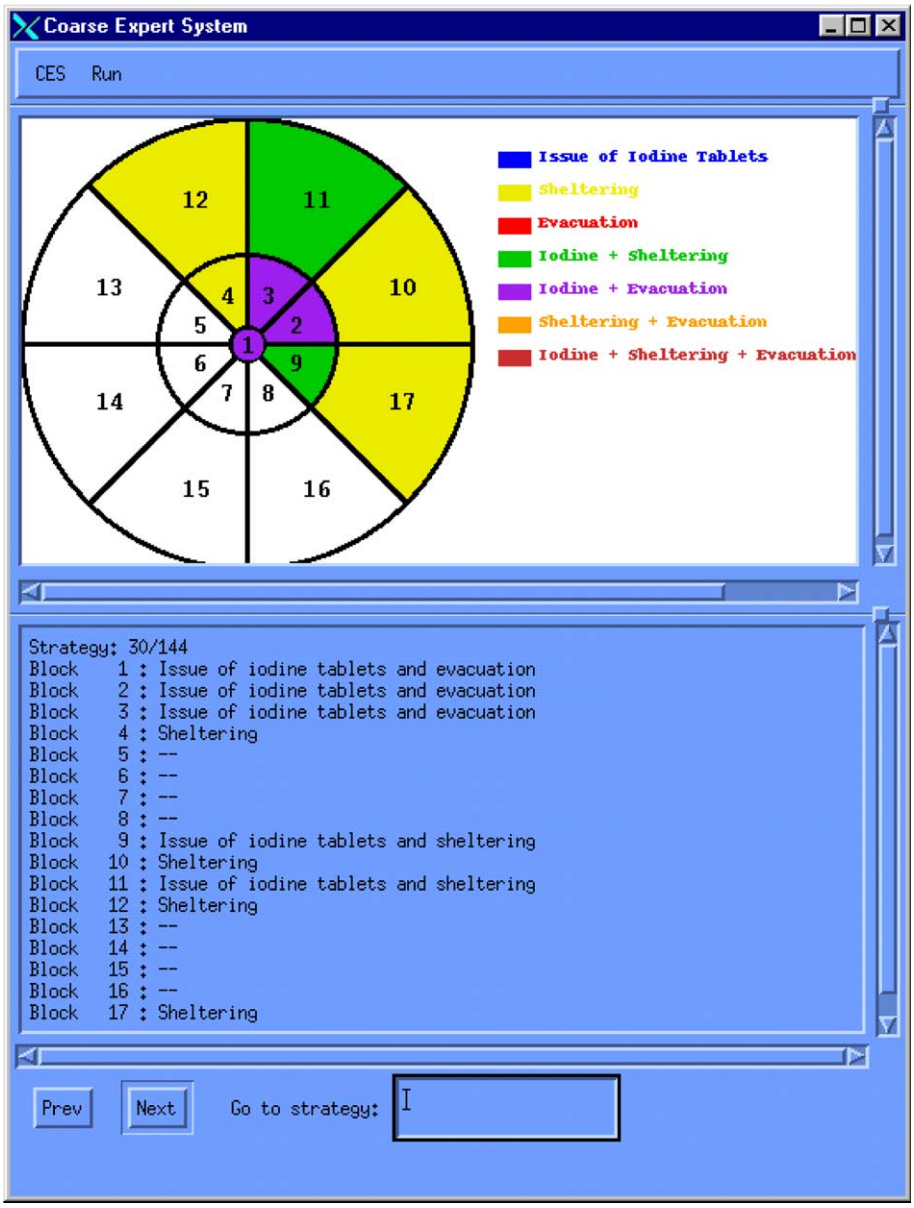

Fig. 3. Coarse expert system output.

less, we have included cost in the model for completeness.

DMs can interact with the user interface of the Ranking Module and select which of the above criteria to consider. Other attributes that might be of interest are the number of thyroid cancers or other related cancers, the technical feasibility of implementing a strategy and other political issues [35]. RODOS modules currently estimate the number of deaths due to cancer but not the number of cancer incidents or the number of thyroid cancers incurred in children. Moreover, they do not give rough estimates of nonquantifiable attributes such as feasibility or psychological effects.

Once the consequences of the strategies over the selected criteria are calculated, the Ranking Module

uses an additive value function to rank the strategies. The ranking process may appear simplistic but the output of the RODOS modules that simulate potential countermeasures is deterministic, which has not allowed us to model uncertainty. Our chosen ranking method is Multi-attribute Value Theory and has been applied in the Ranking Module for several reasons. It replaces a complex decision problem with simple ones, the tradeoffs between attributes and preferences over alternative outcomes are expressed explicitly [48] and the mathematics is relatively simple and easy to understand. It has already been used in nuclear technology applications, especially in siting problems [46], and the International Commission for Radiation Protection (ICRP) recommends its use [42]. The method requires the clear representation of the decision problem through the construction of an attribute tree which eliminates biases in favour of or against some attributes [22] and helps DMs explore a problem [47] and justify their decisions to the public [35].

The user interface of the Ranking Module (Fig. 4) illustrates the decision parameters and the results of the decision analysis. At the top of the user interface, DMs can view the criteria they have selected structured in the form of an attribute tree. In the middle, a histogram illustrates the weights of the criteria. The 10 highest ranked strategies appear at the bottom of the interface (the number of strategies to display is user-defined). The weights of the criteria and the scores of the strategies can be viewed and modified through the interface. When the values of these decision parameters are refined, the list of the top strategies is dynamically updated to reflect changes in the ranking.

The users of the Ranking Module can conduct sensitivity analysis (Fig. 5) to find how robust the top strategies are. They can also view the most

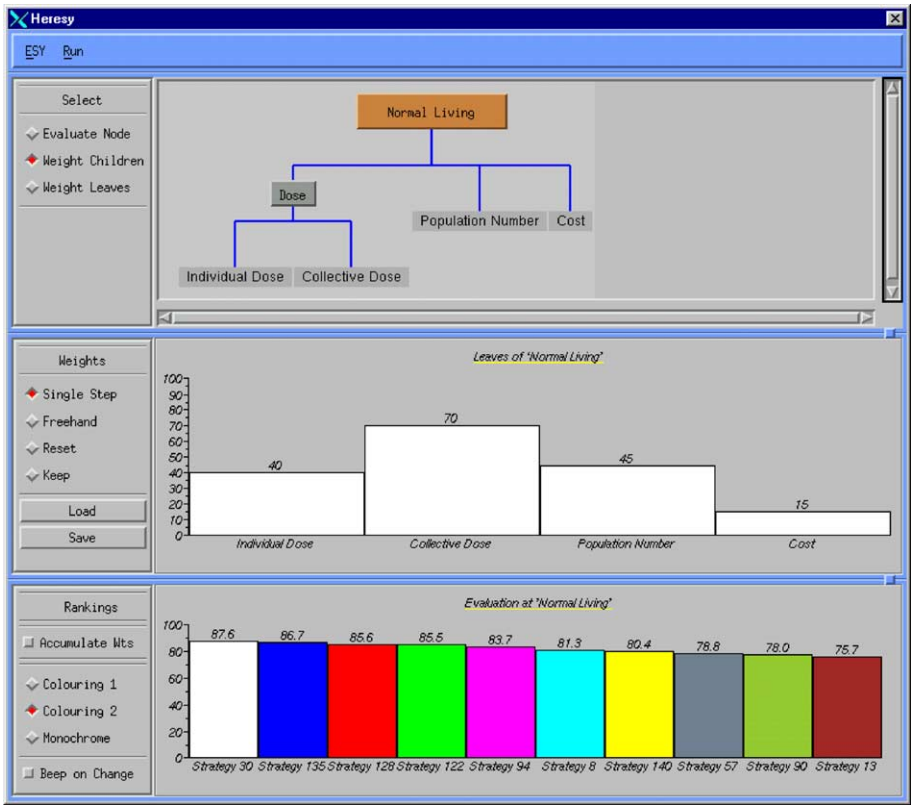

Fig. 4. Ranking module.

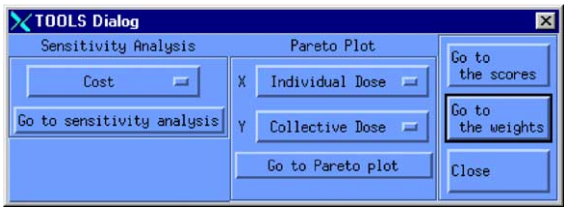

Fig. 5. Other ESY facilities.

efficient or dominant strategies when two criteria (e.g. collective and individual dose) are taken into account through two-dimensional Pareto plots. They may then choose to progress with the decision analysis and obtain explanation reports that interpret the results of the sensitivity analysis as well as highlight the most significant factors in the choice between two alternatives. Alternatively, they may decide to repeat the evaluation process by reformulating the decision problem and refining their preferences. At the end of the analysis, they can choose the strategy they are most comfortable with.

2.4. Fine expert system

Previous studies [91] have established that 'the dogmatic advice of a DSS or expert system is very likely to be rejected if no explanation facilities are provided'. Explanation tools have been shown to influence user perceptions and attitudes such as trust, confidence and satisfaction levels [21]. Explanation facilities improve performance and learning [32] and help a range of users including experienced professionals and novices [56].

In order to add transparency into the ranking process, justify the advice of the Ranking Module and increase the trust and confidence of the DMs in the results of the ranking process, the Fine Expert System—FES—has been developed to generate two natural language reports:

- A comparative report that compares two strategies and interprets the evaluation results, e.g. how much better one alternative is over another, arguments for or against a choice, whether an objective differentiates between two alternatives and what are the most significant factors in the ranking of alternatives.
- A sensitivity analysis report that explains sensitivity analysis graphs shows the effect of changing the weight of an attribute and gives an overall assessment of the decision parameters.

The system employs natural language generation techniques [81] to produce the two reports in English. Its input comprises qualitative information (the attribute tree) and quantitative data (values of decision parameters such as attribute weights and alternative scores). The user issues a command (e.g. 'generate a sensitivity analysis report on the weight of an attribute') through the user interface (Fig. 6). The command is translated into a communicative goal that expresses the purpose of the text to be generated and is posted to the natural language generator. The text planner, which is the first component of the generator, examines the goal and determines what information to communicate to the users and how to structure the information in a coherent way. It has access to a library of text plans and depending on the communicative goal it chooses an appropriate hierarchy of messages. A template-based sentence generator, which is the second component of the system, parses the hierarchy to process its messages and generate text. A set of rules [49] is used to select an appropriate template for

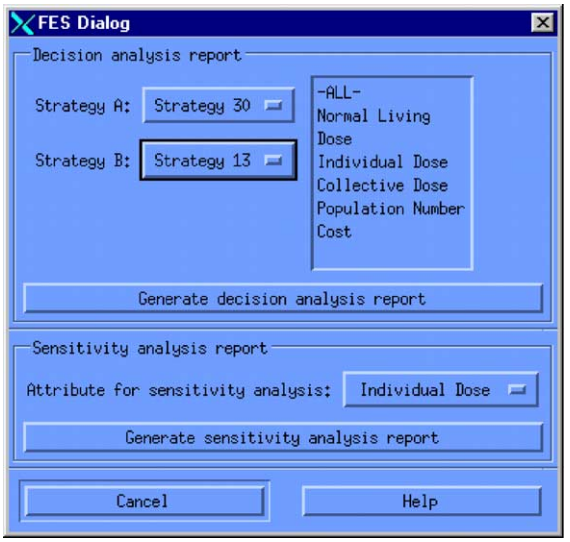

Fig. 6. Fine expert system.

each message (e.g. ⟨Alternative⟩ rates ⟨score⟩ relative to ⟨Objective⟩ on a scale from 0 to 100) and fill in the slots of the template with natural language phrases or quantitative values. The output reports are generated in html format and can be displayed in any Web browser. This has the advantage that the layout of the text is determined during the generation of the sentences.

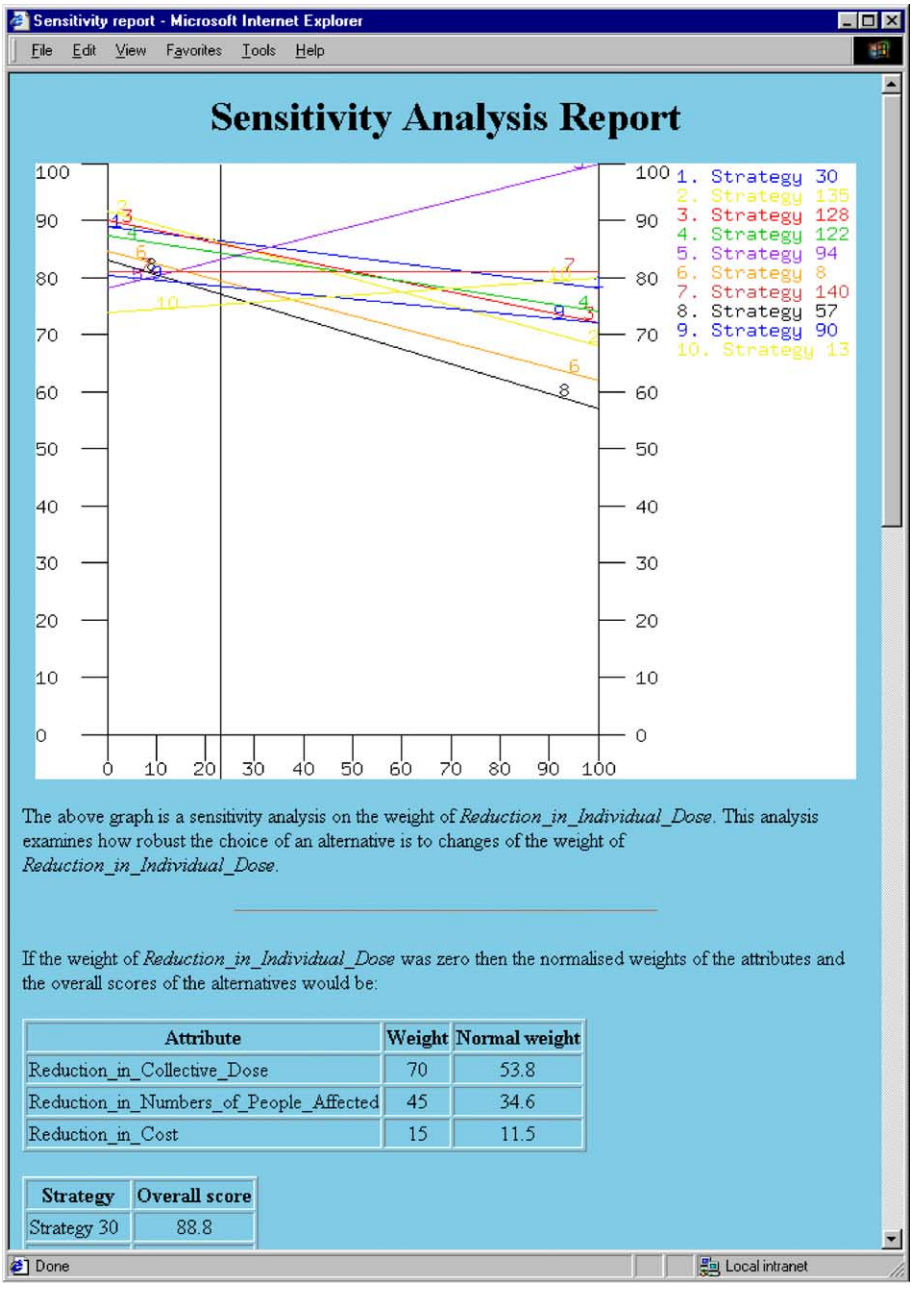

Fig. 7. Sensitivity analysis report.

An extract of a comparative report as generated by the system is as follows:

FES Report

Strategy 30 vs. Strategy 13

Return_to_Normal_Living

Strategy 13 provides slightly lower *Return_to_Normal_Living* than Strategy 30.

- This judgement takes into account the effects of *Reduction_in_Dose*, *Reduction_in_Numbers_of_People_Affected* and *Reduction_in_Cost*.
- While *Reduction_in_Numbers_of_People_Affected* is the main reason to prefer Strategy 13, this is outweighed by considerations of *Reduction_in_Cost*, along with other less important factors, that provide reasons for preferring Strategy 30.

Reduction_in_Dose

Reduction_in_Dose is a factor favouring Strategy 30 over Strategy 13 in determining *Return_to_Normal_Living*, although not a strong one.

- This judgement takes into account the effects of *Reduction_in_Individual_Dose* and *Reduction_in_Collective_Dose*.
- While *Reduction_in_Dose* is very important in general for determining *Return_to_Normal_Living*, the differences between Strategy 30 and Strategy 13 are not very large which makes *Reduction_in_Dose* an insignificant factor.
- Strategy 13 provides slightly lower *Reduction_in_Dose* than Strategy 30. *Reduction_in_Collective_Dose* provides the most important reason.
- *Reduction_in_Dose* accounts for 58.3% of the determination of *Return_to_Normal_Living*.

Reduction_in_Numbers_of_People_Affected

Reduction_in_Numbers_of_People_Affected is a factor favouring Strategy 13 over Strategy 30 in determining *Return_to_Normal_Living*, although not a strong one.

- *Reduction_in_Numbers_of_People_Affected* is not contributing in this particular choice because *Reduction_in_Numbers_of_People_Affected* is not so important in determining *Return_to_Normal_Living* and Strategy 30 does not provide particularly different Reduction_in_Numbers_of_People_Affected from Strategy 13.
- Strategy 30 provides reasonably lower *Reduction_in_Numbers_of_People_Affected* than Strategy 13. Strategy 30 provides very good *Reduction_in_Numbers_of_People_Affected* in the context of all available strategies. Strategy 13 provides very good *Reduction_in_Numbers_of_People_Affected* in the context of all available strategies.
- Strategy 30 rates 90.6 on *Reduction_in_Numbers_of_People_Affected* on a scale from 0 to 100. Strategy 13 rates 96.9 on *Reduction_in_Numbers_of_People_Affected* on a scale from 0 to 100.
- *Reduction_in_Numbers_of_People_Affected* accounts for 31.2% of the determination of *Return_to_Normal_Living*.

. . .

Fig. 7 shows an example of a sensitivity analysis report. An extract of the report is as follows:

Sensitivity Analysis Report

. . .

Strategy 135 gives the highest value of *Reduction_in_Individual_Dose* as long as the weight placed on *Reduction_in_Individual_Dose* is less than 21.9. If the weight is between 21.9 and 32.6 then Strategy 30 is the most preferred alternative. If the weight is higher than 32.6 then Strategy 94 has the

highest value of *Reduction_in_Individual_Dose*.

The percentage on Reduction_in_Individual_Dose can be changed by as much as 1.61% without changing the optimality of Strategy 30.

---

3. Evaluation of DSSs and expert systems

3.1. Verification and validation of DSSs and expert systems

Several verification and validation methods have been suggested in the literature. These can be broadly categorised into three types [1,7,78]:

1. Technical methods that include tests for measuring the technical aspects of the system, validating its components and examining what knowledge sources have been used.
2. Empirical methods that include tests for measuring the performance of the system and its users, e.g. whether the DMs make significantly better or faster decisions or use substantially more information.
3. Subjective methods that include tests for measuring the utility of the system, i.e. whether the system addresses an important problem, how logical and systematic its problem-solving approach is, whether the system meets the needs of its users and how well its interface is designed.

As Adelman [1] notes, technical methods focus on internal correctness (i.e. correctness of the numerical and data procedures) and represent verification methods. Verification tests aim at debugging the logic of a computer program and eliminate any errors in the system. Verification is usually part of the software development process. Empirical methods, on the other hand, focus on validation issues such as external correctness (i.e. the quality of a system). They check how well a system performs its tasks and whether it has improved the performance of the DMs. Finally, subjective methods are mainly employed to validate a system but they can also be used to verify it.

This paper draws from recent literature to classify empirical and subjective methods that can be used to assess DSSs and expert systems. Technical methods are mainly reported in the software development literature and are briefly reviewed.

3.2. Panel-based evaluations

This approach is often adopted for the evaluation of decision support expert systems. Its aim is to prove that a decision tool can lead to decisions that are at least as good as the decisions taken without using the tool. This method has been used in various applications such as divorce settlements [99], new product management [78], therapy planning [38], pavement treatment selection [37] and optimisation of complex engineering systems [50]. The outputs of the expert system, in several predefined case studies, are compared against the recommendations of one or more experts to measure the performance of the system. However, instead of using a third party to evaluate the outputs of both the experts and expert systems, the same experts are often used to assess the results. This encompasses the danger that some experts might systematically rate the results of the system as better or worse relative to human produced output. Even if blind tests are adopted, there is the possibility that some experts will rate themselves more favourably or unfavourably than other experts [8].

3.3. Turing tests

These tests are named after Turing who in 1950 suggested that the intelligence of computer machinery could be assessed using the so-called ‘imitation game’ [96]. The idea is that a computer and a person perform the same tasks and answer several questions. An external interrogator then states if the two could be distinguished based on their responses only. Boritz and Wensley [8] apply a Turing test to evaluate an expert system designed to generate audit plans using case studies in retailing, banking and other application domains.

Turing tests differ from panel-based evaluations in that a third party assesses the performances of both the experts and the expert systems. Statistical methods such as percentage agreement, Kappa analysis and cluster analysis [34] can be used to make this comparison more precise and complete. Any problems that may arise during the statistical analysis such as small samples of data due to shortage of human experts or data which does not follow any general distributions can be overcome using computer intensive statistics techniques—these are techniques that recompute test statistics a large number of times in order to develop a resulting distribution of that test statistic based on the original sample—e.g. enumeration and randomisation [68]. Intelligent systems have also been introduced to automate the comparison of the performance of an intelligent system with the performance of human experts [60].

The case tests used in the panel-based and Turing tests assume that an expert system's performance should match the performance of an expert. However, experts can make mistakes in their judgements or predictions and it might be dangerous to try to emulate their performance. Another problem is finding some suitable test cases to make the comparisons. Selecting the right experts could also prove to be difficult. Finally, if the knowledge base or the inference engine of an expert system changes (e.g. rules, data) then a panel of experts has to be invited again to reassess the expert system.

3.4. Validating the performance of expert systems and models

The performance of knowledge-based systems can be validated using a wide range of methods including benchmarking and sensitivity analysis [66]. Software tools can be developed to automate the process [61]. Several criteria such as consistency, completeness and exactitude (i.e. precision, accuracy and definition) have been suggested in the literature to validate mathematical models [26]. Key requirements including knowledge capture, organisation, formalisation, distribution and application should be identified in the application domain of a DSS prior to its evaluation [63].

3.5. Comparison of a DSS with other systems in the same domain

The performance of a DSS can be measured by comparing it with other DSSs. Objective measures are decided in advance. For example, Richardson [82] compares the ability of a DSS for academic libraries to identify references and the accuracy of its responses relative to other Web-based systems. Subjective measures such as ease of use and user satisfaction have been defined and potential users make the comparisons. In another study [75], a knowledge refinement system is compared to other representative systems in the same field in terms of expressive power, applicability, flexibility and accuracy. Other studies compare the functionalities of a DSS to the features of other distributed DSSs [16].

3.6. Assessing the quality of a decision

Several studies assess the performance of a DSS by measuring the quality of the decisions taken by the DMs after using the system. Sharda et al. [86] examine the effectiveness of DSS-aided DMs relative to DMs without any DSS support by measuring the quality of their decisions in a business-simulated game. They also discuss the results of similar studies, according to which the quality of a decision can be assessed by raters or measured using either subjective factors such as user confidence or objective factors such as decision time, volatility in profit, number of alternatives considered and amount of information requested. Belardo et al. [5] examine the usefulness of incorporating microcomputers into DSSs for nuclear emergency management by calculating the number of correct decisions that the subjects of the study take. They regard a decision as correct when it is taken in accordance with an emergency plan because in that way the decision is expected to lead to fewer casualties and less property damage.

Measuring the quality of a decision based on decision outcomes, e.g. profits, has several disadvantages. Good decisions can lead to bad outcomes such as loss of profits and bad decisions may result in good outcomes. If a DSS gives accurate and complete advice in a specific set of test cases, this does not necessarily mean that its reasoning is sound and

consistent. Moreover, as Todd and Benbasat [92] point out there is a 'black box' relationship between the output of a decision aid and the outcome of the decision taken with the help of the tool. Differences in decision-making styles due to individual characteristics such as cognitive processes or biases have an impact on the interactions of DMs with a DSS [51].

Mackay et al. [55] evaluate the effect of decision aids on the problem-solving process rather than the outcome of the decisions taken by DMs who use these aids. They outline problem solving processes such as problem finding, problem formulation, idea generation and solution identification and they measure the time spent by DMs, the number of activities or actions in each problem-solving process and the total time needed to reach a final decision. Silverman [88] proposes a framework for measuring the performance of expert critiquing systems by calculating the number of utilised cues, i.e. knowledge chunks, lessons learnt and rules used for making a judgement. Personality and cognitive style instruments are used to interpret differences across the DMs' responses. However, studies like this one require extensive analysis and it might be difficult to define the utilised cues.

3.7. Direct assessment

In several studies, the subjects directly assess their own performance and/or the performance of a DSS. For example, self-assessment questions were used in Ref. [44] to check the usefulness and suitability of a DSS, the quality of the arguments provided as well as the productivity, responsiveness, organisation of thoughts, cogency of arguments, learning ability, quality of analysis and confidence of the DMs. The questions were administered before and after training sessions on credit management to experimental groups that used the DSS to reach a conclusion and to control groups that did not use the DSS. In another study [40], subjects reviewed interfaces of information retrieval systems and indicated their level of satisfaction and cognitive load for each design. Metrics can be devised to measure the effectiveness of performing a decision task [12], the effectiveness of guiding the decision process [74] and the efficiency of making accurate decisions [11]. Other studies examine the ability of a DSS to eliminate errors in medical diagnosis [33] and converge opinions [98]. Users can be invited to offer comments and qualitative feedback throughout the development cycle of a system and improve its interface [15]. Once a prototype is developed, the performance of the system can be tested [30] and the attitudes of potential users towards the system can be established [45] often under laboratory conditions [16]. Case studies can be used to test the system in real-life examples and highlight areas for improvement (see for example Refs. [57,97]).

3.8. Multi-criteria decision analysis techniques

Multi-criteria decision analysis techniques can be used to assess one or more DSSs. Evaluation criteria are identified first and structured into hierarchies. They are allocated a weight or are rated with respect to their intensity level depending on the decision analysis technique used. Potential users or experts rate the DSSs relative to each criterion. An overall score is calculated for each system. If this score is low for any particular DSS, then its introduction is doubtful. Adelman [1] proposes a variety of multi-criteria decision analysis methods ranging from cost-benefit analysis to multi-attribute utility theory (MAUT). Caro [10] discusses the use of a decision matrix for the evaluation of emergency management support systems. Gass [29] adopts the analytic hierarchy process (AHP) method to assess computer-based Operations Research models. Bailey and Pearson [2] measure computer user satisfaction by summing up the user's weighted reactions to a set of factors.

3.9. Questionnaire

Questionnaires are a popular way of determining a user's attitude towards a DSS. They have been used in a variety of applications such as marketing [52], laser safety [14], printed wiring board assembly [3], military command and control [53], strategic decision making [13] and assessment of user interfaces [41].

Questionnaires are usually used to measure perceived usefulness and perceived ease of use. Davis' technology acceptance model [19], which provides a framework for measuring beliefs and predicting a future behavior, suggests that perceived usefulness and perceived ease of use are primary factors for the prediction of computer user acceptance behavior. Other studies have identified more

evaluation criteria. Bailey and Pearson [2] outline 39 factors affecting user satisfaction. Sharma et al. [87] explain what features and characteristics are most critical for the successful implementation of a system. Finlay and Forgani [24] classify a large number of success factors that are important in the development of DSSs. Caro [10] outlines criteria for the selection of an expert system for emergency management training. Forgionne [27] proposes measures for assessing the decision process and outcome in health care applications. Grabowski and Sanborn [31] propose a set of criteria for the evaluation of intelligent real-time systems. Adelman [1] discusses at length what factors to consider when evaluating a DSS.

A Likert-type questionnaire allows the users of a DSS to rate several aspects of a DSS's performance on a scale but it does not allow them to justify their responses. In order to identify what the users perceive to be the strengths and weaknesses of a system, several instruments have been used in the literature, e.g.

- open-ended questionnaires often combined with an interview [1];
- content analysis [14] which is a technique for identifying specified characteristics of messages, texts and communications like 'favourable' and 'adverse' within interview scripts; and,
- protocol analysis [88] where a user interacts with a system and all her keystrokes are saved in a trace file as well as her thoughts, frustrations and positive effects.

3.10. An organisational perspective

Apart from evaluating the interface between a user and a system, DSS developers should also evaluate the interface between the DSS and the organisation where it is installed as well as the interface between the DSS and the organisational environment [1]. A DSS can be considered to be successful when it satisfies not only the needs of the user but also organisational objectives and structures as well as the demands of the organisation's environment which can affect the performance of the DSS. There have been several studies on the evaluation of DSSs from an organisational perspective. Lu [54] proposes a framework for evaluating group support systems. The framework analyses the relations between technology, problems, people and organisational structure, culture and environment. It also evaluates the design of a group support system, the people's acceptance of the technology, and any changes in the organisational culture and setting. Sharma et al. [87] discuss success factors that are most critical in the implementation of expert system decision aids and how the understanding of the associations between these factors is important in the evaluation of such systems.

4. Technical verification and performance validation of the ESY

A variety of methods have been employed to assess the technical aspects of the ESY and its performance. More precisely, we have adopted the following approaches.

4.1. Assessment of the methods employed

We have examined the appropriateness of the methods used in the ESY (e.g. constraint satisfaction techniques to identify feasible alternatives, multi-attribute value theory to evaluate alternatives and natural language generation to explain the system's recommendations) and we have justified their use [70] (see also Section 2).

4.2. Software verification and testing methods

In order to test and verify the ESY code, we used static testing methods (e.g. quality assurance facilities and static analysers [20]) that examine a system's design and software without executing its code as well as dynamic testing methods that execute the system's code using different sets of data.

4.3. Evaluation of the knowledge base

The ESY contains components that codify knowledge and inferences, i.e. rules to solve specific problems. We have performed several checks and consulted experts to make sure that the knowledge base of the ESY is accurate, consistent and

complete and that the reasoning of the system is sound.

4.4. Documentation

Since comprehensive documentation is one of the marks of a quality analysis [28], we have produced detailed documentation for the ESY to describe its user interface and functionalities as well as the ESY interfaces with other RODOS components [70].

4.5. Comparison to other DSSs

We have compared the ESY to other evaluation systems for nuclear emergencies [72] including MOIRA [43], CMDSS [85] and a spreadsheet-based DSS [76]. Unlike other systems, the ESY can be adapted to support decision-making throughout all the phases of a radiation accident. It assists DMs in designing and exploring strategies and explains its recommendation. This helps DMs to identify key factors in the ranking of alternatives, gain insight into the decision process and take a decision based upon understanding.

4.6. Direct assessment

We have demonstrated the ESY at several venues across Europe and had it directly assessed by experts. The objectives of these demonstrations and assessments were the following:

1. To discover any omissions or errors in the knowledge base of the ESY.
2. To examine the reasoning of the system.
3. To discuss the performance of the ESY.
4. To explore national radiological guidelines in other European countries and consider whether to adapt the ESY to meet these guidelines.
5. To identify the needs of the DMs.

5. Subjective assessment of the ESY

5.1. Criteria

Measuring the perceived utility of the ESY required the evaluation of the system from the perspective of its users. A two-part questionnaire was designed to elicit data from the subjects about the system's strengths and weaknesses. Even though the main aim was to measure the utility of the ESY, other criteria that could contribute to the overall utility of the system were also identified. These criteria (Fig. 8) and their definitions are given below (the definitions are taken from Bailey and Pearson [2]):

1. *Perceived utility*: the user's judgement about the relevant balance between the cost and the considered usefulness of the DSS.
2. *Relevance*: the degree of congruence between what the user wants or requires and what the DSS provides.
3. *Understanding of the system*: the degree of comprehension that a user has about the system or services that are provided.
4. *Completeness*: the comprehensiveness of the output information content.
5. *Format of output*: the layout design and display of the output contents.
6. *Volume of output*: the amount of the information given to a user.
7. *Ease of use*: the amount of effort required by the user to take advantage of the tools provided by the system.
8. *Ease of learning*: the potential of a system to require minimal effort in learning how to use it.

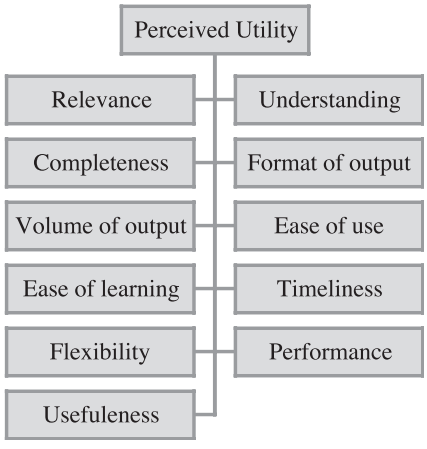

Fig. 8. Hierarchy of evaluation criteria.

Table 1
Subject characteristics (n=21)

Level of expertise	Mean score	Standard deviation
Emergency management	4.81	1.86
Computer tools	5.00	1.18
Decision-aiding tools	5.05	1.35

9. *Timeliness*: the availability of the output information at a suitable time.
10. *Flexibility/adaptability of the system*: the capacity of the system to change or to adjust in response to new conditions, demands, or circumstances.
11. *Performance*: the ability of a DSS to help a DM accomplish a task more effectively.
12. *Usefulness*: the extent to which an application contributes to the enhancement of the user's performance.

5.2. Subjects

Twenty-one subjects took part in the evaluation of the ESY. The subjects ranged from scientists and nuclear plant operators to planning officers and advisors in emergencies. They came from European Union, Eastern European and Former Soviet Union countries.

The questionnaire examined the background of the subjects to establish their level of expertise or experience in different fields (e.g. nuclear emergencies, computers and decision-aiding tools). More precisely, the subjects had to give a score—on a scale from 1 (not experienced at all) to 7 (very experienced)—against the following statements:

Statement 1 Please state your level of expertise in nuclear emergencies.
Statement 2 Please state your level of experience and training with computers.
Statement 3 Please state your level of experience and training with decision aids.

Ten of the subjects had expertise in nuclear emergencies and radiation protection issues (rated 6 or 7 on statement 1) and only two had little expertise (rated 1 or 2 on statement 1). Fifteen subjects were quite familiar with computers (rated more than 4 on statement 2). Similarly, 15 subjects were quite familiar with decision aids (rated more than 4 on statement 3). The mean scores of the subject's responses can be found in Table 1.

5.3. Questionnaire A

In order to measure the criteria defined above, we used a seven point Likert-type scale. In the first part of the questionnaire, the subjects had to state how much they agreed or disagreed with a statement on a scale from 1 ('Very strongly disagree') to 7 ('Very strongly agree') with 4 as the midpoint ("Indifferent"). There were 19 statements (Q1–Q19) in total. A complete list of statements is given in Appendix A. The responses of the subjects are summarised in Table 2. The first two columns of Table 2 show which statements correspond to each of the evaluation criteria. Because we regarded the perceived utility and the relevance of the information conveyed as the two most important criteria in our study, we used more than one statement to measure them, which allowed us to conduct a reliability test (see Appendix B). The statements that corresponded to the same criterion were not sequential in the questionnaire to prevent the order of the statements from affecting the subjects' scoring. Statements Q2, Q3 and Q4 measured the utility of ESY components that contributed to the overall utility of the system.

Two subjects reviewed the questionnaire before its distribution to make sure that none of the statements was ambiguous or difficult to understand.

Table 2
Evaluation criteria and average ratings (max 7; min 1)

Criteria	Questions	Mean	Standard deviation
Perceived utility	Q1, Q5, Q19	5.50	1.32
Relevance	Q6, Q11, Q16	5.06	1.33
Understanding	Q7	4.76	1.44
Completeness	Q8	5.30	1.08
Format of output	Q9	5.30	1.21
Volume of output	Q10	5.43	1.12
Ease of use	Q12	4.48	2.01
Ease of learning	Q13	5.05	1.65
Timeliness	Q14	4.95	1.35
Flexibility	Q15	5.10	1.13
Performance	Q17	5.57	1.28
Usefulness	Q18	5.00	1.68

5.4. Results of Questionnaire A

The mean scores of the responses of the subjects over the statements (Q1–Q19) range between 4.48 and 5.67. As it can be seen in Table 2, the mean scores of eight (out of 12) criteria are greater than 5.0 on a scale from 1 to 7. The mean scores of the other criteria are between 4.48 and 5.0, i.e. higher than the midpoint (4). This suggests that the ESY met the evaluation criteria we set at the beginning of the study. The criterion that received the lowest score was 'ease of use'. Some subjects felt that the ESY was somewhat difficult to use and that they would need the assistance of a technical person to operate it. However, the subjects rated the criterion 'ease of learning' with a higher score (5.05). The criterion that received the highest mean score (5.57) was 'performance'.

As shown in Table 3, the Coarse Expert System facilities (generation of alternatives) rated 5.57, the Fine Expert System facilities (justification of the system's advice) 5.33 and the sensitivity analysis facilities were given a mean score of 5.45. The other statements that directly referred to the overall utility of the ESY received the following mean scores: Q1 (5.52), Q5 (5.33) and Q19 (5.67). While the mean score of the statements that directly referred to the overall utility (Q1, Q2, Q3, Q4, Q5, Q19) was 5.48, the mean score of the remaining 13 statements was 5.09, which indicates that the utility measurements were precise. Similarly, the measurements on the relevance of the ESY outputs and functionalities are Q6 (4.95), Q11 (4.86) and Q16 (5.37). It is noted that statement Q16 refers to the relevance of the ESY output to RODOS (the majority of the respondents were more familiar with RODOS rather than a general evaluation setting).

The reliability of Questionnaire A is examined in Appendix B. This is followed by the validation of Questionnaire A in Appendix C.

Table 3
ESY facilities (max 7; min 1)

Facilities	Mean	Standard deviation
Automated generation of alternatives	5.57	1.12
Automated generation of explanations	5.33	1.27
Sensitivity analysis	5.45	1.23

5.5. Questionnaire B

Questionnaire A gave the opportunity to the subjects to express how much they liked or disliked the ESY. However, it did not allow them to justify their opinions. Another questionnaire was therefore constructed in which the subjects were able to highlight problem areas of the interface and of their interaction with the system and write how useful the ESY was in their own words. The questionnaire was open-ended and therefore analogous to an interview [1] asking questions such as:

1. What did you dislike and/or find most restrictive or ineffective about the ESY? Why?
2. What do you think would be the potential of the ESY in the event of a radiation accident? Why?

The complete list of the questions asked is given in Appendix D. It should be noted that apart from the 21 subjects who answered Questionnaire A, some additional subjects answered Questionnaire B or commented on the ESY during demonstrations.

5.6. Results of Questionnaire B

The subjects suggested several improvements on the ESY. These improvements can be grouped into the following categories.

5.6.1. Modelling issues

The user interface of the Coarse Expert System illustrates how the area around the nuclear plant can be divided into emergency planning sectors and zones. Because the number of sectors or zones for emergency planning purposes varies across Europe, some subjects expressed concerns. After the ESY assessment, it became clear that we either needed to build different versions of the Coarse Expert System tailored to the national legislation or needs of DMs in each country or make the Coarse Expert System more adaptable so that the number of sectors or zones is a user-defined variable. We also decided to build a version tailored to the needs of Finnish DMs to demonstrate the applicability of the system in different contexts (emergency planning in Finland is based on municipalities rather than zones and sectors). We have yet to satisfy a user's requirement to allow nuclear site

administrators and other users to define their own constraints based on national or local conditions and regulations.

5.6.2. User interface

There were many constructive comments on the ESY interface. For example, some subjects pointed out that they would prefer to see or be able to set a name for each alternative instead of an index number. Some inconsistencies were detected on the menu bar items through all the ESY windows. Another criticism was that the ESY graphical displays should be more in line with the windows of other RODOS modules (e.g. menu bars).

On a positive side, many subjects felt that the interface was user-friendly and that the system responded very fast to requests. The alternative strategies were clearly shown on the user interface of the Coarse Expert System (Fig. 3). A subject remarked how good the colours looked. Another subject however noted that, knowing the abilities of many DMs, the ESY should be more 'flashy'. It is remarkable how different the expectations and perceptions of the interface were.

5.6.3. Positive aspects of the ESY

The subjects stated the following about the ESY:

- It provides comprehensive decision support (*comprehensiveness*).
- It helps DMs think about decisions in a structured and systematic way (*learning and support*).
- It allows DMs to thoroughly explore alternatives and gain a better understanding of the decision problem (*understanding of decision content*).
- It shows the effect of varying decision parameters (*technical capability*).
- It makes the decision-making process more transparent (*transparency*).
- It has a clear and understandable structure (*understanding of DSS*).

The comments of the subjects are particularly important because they highlight factors that can potentially contribute to the successful implementation of a system. Providing a wide range of facilities and comprehensive output are desirable aspects of a DSS [2]. A learning and support environment that allows DMs to learn about the decision domain in a structured way is very important in improving decision effectiveness [24,44,74,78]. The ability of a DSS to generate a complete set of alternatives [63] and encourage DMs to explore a large number of alternatives [86] are important features. Technical capability, in particular, is a key success factor [1,24,78]. Adding transparency to the decision process improves public understanding and acceptance [73]. Finally, understanding the services and functionalities of a DSS has been found to contribute to its successful adoption [2,24].

5.6.4. Suggestions for improvement in the ESY

A subject felt that the ESY's comprehensive decision support was disadvantageous. The subject stated that the ESY supported a wide range of decisions and said that different DMs should be involved in different decision-making aspects. Different levels of decision support should be provided depending on the background of the DMs.

The ESY was found complex by a subject; either a technical team should be available during a nuclear emergency to operate the system or DMs should try to learn how to use the ESY to avoid any misunderstandings about the functionalities of the system during its operation. It seems likely though that technical people will be available for the operation of the RODOS system during a nuclear emergency.

A subject stated that he/she would prefer to discuss the results (e.g. measurements) with an expert who would interpret these results against intervention levels rather than interpreting the outputs of computer programs. Apparently, the subject did not take into account the real possibility of having very scarce and erroneous measurements in the early phases of a nuclear emergency. We acknowledge, however, that some people feel more confident and satisfied when interacting with human beings rather than computer programs. It is expected that the ESY will be used with the help of a technical person and that a decision analyst will assist in the decision-making process.

The explanation facilities of the ESY were well received. Some subjects, however, preferred examining the ESY results in tabular format and interpreting ESY graphs rather than read text reports. However, novices and subjects who were not familiar with

decision analysis issues found the text reports to be useful. Many subjects suggested extending the explanation facilities to describe the best alternative strategies (e.g. which areas should be evacuated, number of people involved) and their consequences (e.g. health effects, economic losses).

Another finding was that the sensitivity analysis graphs were confusing and difficult to interpret for several subjects. Only a few of them were familiar with such plots. The Fine Expert System produces a report that interprets sensitivity analysis graphs, which was found to be a useful functionality.

5.6.5. Potential of the ESY

The subjects felt that the ESY could be very useful in the medium phases of a nuclear emergency where the DMs would be able to explore different alternatives without time pressures. Some subjects felt that the use of the ESY would be limited in the early phases of an accident because of time constraints. A subject however strongly believed that the use of the ESY in the early phases would be very beneficial because the Coarse Expert System automatically produces feasible strategies whereas DMs find it very difficult to come up with some good alternatives for evaluation. The ESY would be more useful to those DMs who were familiar with it and understood its functionalities.

The vast majority of the subjects believed that the ESY could be a valuable tool for training purposes. More precisely, DMs could use the ESY as a training tool to:

- identify alternatives in different nuclear accident scenarios;
- explore the effect of changing the values of decision parameters in the ranking of the strategies;
- consider factors such as social effects that are difficult to grasp;
- practice how to take decisions under different circumstances.

According to the subjects, other areas in which the ESY could be used include emergencies in general, environmental decision problems and decision problems with a large number of evaluation criteria or alternatives.

6. Discussion and lessons learned

Throughout this evaluation study, we encountered challenges and had to address technical and cultural issues. In this section, we highlight some of our experiences in designing, developing, operating and assessing the ESY.

There is a wide range of methods and tools that can be used to assess the technical aspects of an intelligent DSS and validate its components. DSS builders should take into account the feasibility of conducting validation tests [25]. In this application, we mainly used methods we were familiar with or tools that were readily available to us. A more comprehensive survey would have allowed us to assess the suitability of the methods/tools used to assess the ESY. A meta-evaluation process could be instantiated to control the activities of the evaluation process.

Devising an evaluation strategy during the development cycle of a DSS entails choosing appropriate methods for building the DSS components, which is a critical and often difficult decision [1]. When developing intelligent DSSs in particular, different techniques can be used to incorporate intelligence into the system. The selection of a suitable approach can be elaborate and the system developers might not be familiar with a particular method. In this work, we occasionally had to experiment with different approaches and opting for a method was often a process of trial and error.

Building the ESY components and assessing the technical aspects of the system was relatively easier than obtaining subjective opinions. In order to test the technical characteristics of the ESY, we conducted experiments, observed the system in well-defined test scenarios and measured its performance against several objective criteria such as robustness and output accuracy. During the subjective evaluation, however, it was not clear how the subjects perceived the statements and questions of the questionnaire. Some subjects who participated in the subjective evaluation of the ESY found it difficult to articulate their thoughts and score the system.

Perhaps, one of the most challenging evaluation tasks is to assess whether a DSS improves the performance of its users. It is feasible to measure the efficiency of decision making (i.e. reduction in

decision time and costs) but as we discussed in Section 3.6, it is not advisable to assess the effectiveness of taking a decision based on decision outcomes. Criteria that could be used to measure the quality of a decision and therefore effectiveness of decision making are the number of alternatives considered [86], the amount of information used (or number of requests for information) [88] and the confidence or satisfaction of the users in the results of the system [74].

After collecting the responses of several subjects to our subjective evaluation, we realised that we had not included 'trust' (i.e. the degree of belief or confidence in the ESY results) as another criterion in our study, which is an important criterion in the evaluation of DSSs [2]. Therefore, the list of criteria used to assess the system was not complete. It was then difficult to collect the opinions of the subjects on this particular aspect. Checking the completeness of the evaluation framework is therefore very important when designing questionnaires [1].

Apart from examining the evaluation criteria used to assess a system, a thorough check of a questionnaire prior to its distribution is equally important. As shown in Appendix A, the statements of questionnaire A were all positively framed. A score of 7 always indicated that the subject strongly agreed with a statement and therefore the rating of the associated criterion was high. We cannot conclude whether all subjects read carefully all statements before giving their judgments. Perhaps, some subjects consistently scored the statements highly or lowly based on their overall assessment of the system. It is better practice to use questions or a mixture of positively and negatively framed statements when designing questionnaires [1].

Developing a DSS for use across Europe creates the need for addressing cultural issues in the design and operation of the system. Several empirical studies discuss the topic but there is clearly the need to further examine differences in the use of DSSs between cultures [94]. Apart from cultural concerns, in some application domains (e.g. crisis management) there are different regulations and policies in place regarding emergency response. For example, in our study we tried to incorporate intelligence into the system by defining rules and constraints that allow the automatic generation of feasible alternative strategies. Such rules vary across Europe and we need to understand the regulations that are in force in each European country, prior to the installation of the DSS. Encouraging users to contribute rules and default values can potentially increase their satisfaction and confidence levels.

Explanation facilities have been shown to influence confidence and satisfaction levels [21]. The generation of explanations facilitates learning and improves performance [32]. New frameworks of decision support suggest the use of natural language for presenting the results of a DSS and justifying the models used [62]. The ESY outputs reports to communicate its recommendation in a natural language form. A dilemma encountered was to choose between the degree of intuition that the ESY explanations exhibited and the formality of the explanation. The richer the vocabulary the higher the risk of describing the same concept with substantially different expressions which might increase the ambiguity of the report. Care was taken to ensure that the vocabulary used in the generated reports did not contain any ambiguous words or concepts that may be misinterpreted by some Europeans. Empirical studies could be conducted to determine the content of the explanations provided and the way users submit input and perceive output in natural language form.

Most subjects liked the ESY but they concluded that they needed the help of a technical person to use it. Based on the qualitative feedback we received, we have concluded that it is not always beneficial to build a system that provides a variety of functionalities in order to please as many users as possible. A system that provides a wide range of functions is less restrictive but if it is complex and difficult to use, its users might reject it [74]. Ease of use is a key success factor [24] and is linked to increased confidence levels and usage [41]. More studies are needed to highlight the need to achieve a balance between simplicity and completeness of functionalities provided.

Another challenge in our study was getting hold of prospective users to evaluate the ESY throughout its development. This was mainly because of the application domain (i.e. crisis management). Potential DMs included politicians and emergency management officers who were not easily accessible. Even identifying the users of the system was not always

straightforward because of intermediaries. For example, in some emergency planning exercises we conducted [4], the users of the DSS modules were scientists or operators. The outputs of the modules were then printed and passed to the DMs who were located in another emergency management room.

As it was expected, the evaluation of the ESY revealed that its potential users had different backgrounds and requirements. DSS developers have to recognise that users have diverse experiences, training, beliefs and preferences, which often makes them understand and operate a system in a different way [54,55]. Therefore, it may be beneficial to customise a DSS so that it addresses the needs of individual DMs or to provide levels of decision support to assist a range of DMs (from novices to experts). Another approach might be to build a system in a narrow domain for a specific group of users that share common experiences and have uniform needs.

In our study, some DMs stated that they were reluctant to use the results of a DSS to take decisions. In a real emergency, they would rely on the availability of measurement data that indicate the scale of an accident and the help of advisors. This category of DMs usually trusts experts and is reluctant to consider the advice of a system. Demonstrating a DSS in a variety of test scenarios and training potential DMs are necessary steps toward improving the system's acceptability and increasing the confidence and trust of the users in the results of the system.

The vast majority of the subjects we interviewed believed that the ESY would be useful as a training tool in emergency exercises. Even those subjects, who had reservations about using the system in a real accident, believed that a DSS could improve their decision-making skills if used for training purposes. Emergency management training systems can help novice DMs learn about the content and process of decision making [10].

Taking a decision in the event of radiation accident is very complex; DMs are faced with a plethora of data, conflicting objectives and uncertainty. A DSS can considerably reduce the amount of time required to process data, assess the radiological situation and explore alternative strategies. As shown in a previous elicitation exercise [4], the main advantage of using the ESY is that DMs directly explore alternatives and compare their scores. On the contrary, when they did not use the ESY they had to process a large amount of information about the radiation accident and compare graphs for different types of dose and for different countermeasures, which made it difficult for them to give an overall assessment of the decision problem and choose an efficient alternative.

The ESY received more positive feedback when it was linked to other prediction systems and its input was automatically generated by other models. DSSs that are fully integrated into the operational process might be perceived to be more useful than stand-alone systems [2,10,95].

Building an intelligent DSS is a long process. Blair et al. [6] reviewed 13 intelligent DSSs whose mean development time was 2.6 years. The ESY took about 4 years to develop and it was difficult to have it assessed any time a new component or functionality was added. The system is currently under further development taking into account the results of the assessment.

7. Conclusions

This paper discusses the evaluation of intelligent DSSs. Various methods can be employed including panel-based evaluation, Turing tests, performance validation techniques, comparisons with other DSSs, direct assessment, multi-criteria decision analysis techniques and questionnaires.

We have developed an intelligent DSS to support decision making in nuclear emergencies. The system, called ESY, undertakes tasks such as generating and evaluating alternatives and justifies its recommendation in a natural language form. It guides DMs (i.e. scientists, health officials, emergency planning officers) through the emergency management process.

We have devised a strategy to evaluate the ESY that involves the following three levels:

Technical verification. We checked the technical aspects of the system to find out how well it was built. The appropriateness of the approaches used in the ESY components was established. Static and dynamic testing methods were used during

the development of the ESY to ensure that the code was well written. Even though we did not use any sophisticated software tools to test the completeness and accuracy of the codified knowledge in the ESY, we had the system inspected by radiation protection experts. Different types of documentation have been written to describe the operations and interfaces of the ESY.

Performance validation. We have tested the ESY to establish how well it performs its tasks, how complete its knowledge base is and whether its advice is sound. We demonstrated the ESY in several venues across Europe (EU, Eastern and Former Soviet Union countries) and received positive comments about its performance. Because of time pressures and lack of resources, we have not been able to investigate whether the use of the ESY increases the quality of the decisions taken. However, the results of the subjective evaluation of the ESY indicated that potential users believe that the ESY can improve their performance in executing tasks such as generation and evaluation of alternatives.

Subjective assessment. We measured the utility of the system to find out how the users perceive the ESY, whether the system fits to the users' needs and how well its interface is designed. We identified several evaluation criteria, such as usefulness, relevance and completeness of the output. Twenty-one potential ESY users answered two questionnaires. The results of the questionnaires indicate that potential users believe that the system is useful and provides a structured and well-organised approach for the evaluation of strategies in nuclear emergencies. Their comments are very encouraging and constructive.

Acknowledgements

We would like to thank all our colleagues from the RODOS project for their valuable comments and discussions. Special thanks are due to all those people who assessed the ESY. This work was funded by the Commission of the European Communities (Contract: F14P-CT95-0007). The views expressed in this paper are those of the authors and do not necessarily reflect those of the RODOS project.

Appendix A. Questionnaire A (statements)

Q1. The ESY offers a structured and well-organised approach to evaluate strategies in nuclear emergencies.
Q2. The ESY's ability to generate feasible strategies is helpful.
Q3. The ESY's explanation facilities are beneficial.
Q4. The ESY's provision of sensitivity analysis tools is advantageous.
Q5. My attitude towards the system is very positive.
Q6. ESY provides what is needed.
Q7. My interaction with the ESY is clear and understandable.
Q8. The ESY provides sufficient information.
Q9. The layout of the information displayed is straightforward.
Q10. The amount of information presented is reasonable.
Q11. All the information provided by the ESY is useful and relevant to an evaluation task.
Q12. I think that I would *not* need the support of a technical person to be able to use the system.
Q13. I suppose that most people would learn to use this system very quickly if used often.
Q14. The right information becomes available at the right time.
Q15. I believe that the system can adjust easily to different nuclear accident scenarios or new conditions.
Q16. The ESY provides facilities and information required by an evaluation module in RODOS.
Q17. The ESY enables a decision maker to accomplish a task more effectively.
Q18. If I had to take a decision in a nuclear emergency, I would need the information provided by the ESY.
Q19. Overall, I consider the system to be useful for the evaluation of strategies in nuclear emergencies.

Appendix B. Reliability of Questionnaire A

Reliability is the extent to which a test produces consistent results when administered under similar

conditions [36]. Perhaps, the most effective way to check the reliability of a measurement instrument or test (e.g. Questionnaire A) is to administer the test to the same subjects twice in order to find out whether their responses remain stable over time (test–retest reliability). A high correlation between the two administrations indicates whether the test is reliable. Another reliability check is to administer two parallel tests to measure the same variable and see whether the results of the two tests are correlated.

Because it was not feasible to administer Questionnaire A twice or construct two questionnaires to measure the utility of the ESY, we decided to measure the internal consistency of the questionnaire instead. While test–retest and parallel tests measure the degree to which the responses to a measurement instrument (e.g. questionnaire) are similar at different times, internal consistency indicates the degree to which the items in a measurement instrument correlate with each other.

There are several metrics to measure the internal consistency of measurement instruments. Different approaches to assessing the reliability of a questionnaire are reported in the literature on DSS evaluation, e.g. split-halves method [1], reliability coefficient [2] and Cronbach's alpha [78]. We have not used the split-halves method because it assumes that two questions are being asked for each evaluation criterion. The reliability coefficient in Bailey and Pearson [2] uses a measurement error that is difficult to calculate. Therefore, we have chosen to assess the reliability of Questionnaire A by calculating the Cronbach's alpha metric whose use is the norm in reliability assessment [83].

Cronbach's alpha or α [17] has been used to measure the correlation among the statements Q1–Q19. The higher the correlation the higher the value of Cronbach's alpha. A high correlation indicates that high (or low) scores on one statement are associated with high (or low) scores on another statement. Since statements Q1–Q19 and their corresponding criteria measure the utility of the ESY, the responses to them should be positively correlated. If not, then we cannot expect the responses to be correlated with any other responses to additional criteria and statements.

Cronbach's alpha has several interpretations [64]. For example, it can be viewed as the correlation between Questionnaire A and all other questionnaires that contain the same number of statements (i.e. 19) that could be constructed from a hypothetical universe of statements that measure utility. It can also be viewed as the squared correlation between the score the ESY receives from a person (the observed score) and the score the ESY would have received if the person would be questioned on all the utility statements in the universe. If the reliability model is violated, Cronbach's alpha takes negative values. Otherwise, its values range from 0 to 1, with 1 indicating that a test or measurement instrument is perfectly reliable.

Cronbach's alpha for Questionnaire A was 0.9684. Nunnaly [65] states that the Cronbach's alpha of a scale should be greater than 0.70 for items to be used together as a scale. Questionnaire A is therefore reliable. If we remove any of the statements (Q1–Q19) and its responses from the questionnaire, we observe a small decrease in the value of Cronbach's alpha (the value is between 0.9656 and 0.9699). Thus, we need to keep all the statements and their corresponding criteria in our analysis.

Appendix C. Validity of Questionnaire A

According to Carmines and Zeller [9], validity is the extent to which a measurement instrument measured what it was supposed to measure. Reliability is a necessary but not a sufficient condition for validity [65]. The presence of measurement error decreases the validity of a measurement instrument but even if there is no measurement error at all there is no guarantee of validity. For example, a test which measures how much the DMs liked the colours used in the graphic outputs of the ESY may be quite reliable but it is a poor indicator of the ESY's utility and has therefore poor validity.

It should be noted that validation refers to the results of a measurement instrument and not to the instrument itself. As Cronbach ([18], p. 447) states: "One validates, not a test, but an interpretation of data arising from a specified procedure". This means that the validation of Questionnaire A is not about validating the properties of the questionnaire but it

is about making sure that the conclusions derived from the questionnaire are valid. We have tested three types of validation:

- Predictive or criterion-related validity.
- Face or content validity.
- Construct validity.

Nunnaly [65] notes that predictive validity "is at issue when the purpose is to use an instrument to estimate some important form of behavior that is external to the measuring instrument itself". We assessed this type of validity to show that Questionnaire A accurately predicted how potential ESY users perceived the utility of the system. In order to measure the predictive validity of Questionnaire A, we followed the approach used in Bailey and Pearson [2] and Adelman [1]. We compared the global assessments of the subjects on the utility of the ESY with their responses on the evaluation criteria (e.g. usefulness, relevance, and performance). More precisely, for each subject we compared the mean scores of her responses to the six statements (Q1–Q5, Q19) that directly refer to the overall utility of the ESY with their scores on the remaining statements (Q6–Q18) that correspond to the evaluation criteria. The correlation (Pearson correlation coefficient) for the 21 subjects was 0.83 ($p \ll 0.01$, df=19; df—degrees of freedom).

Face validity depends on the extent to which an empirical measurement reflects a specific domain content [9]. Before we constructed Questionnaire A, we searched thoroughly the literature on evaluation of DSSs and expert systems. We then identified different dimensions of the utility of a system (e.g. usefulness, performance) and we tried to construct statements that reflected these dimensions. Even though we tried to include as many evaluation criteria as possible, we decided to exclude some factors such as accuracy, compatibility and robustness because we felt that it would be difficult for the subjects to assess them. These factors were measured, however, during the technical and empirical evaluation of the ESY.

Construct validity is concerned with the extent to which a particular measurement instrument relates to other instruments that are consistent with theoretically derived hypotheses about the concepts that are being measured [9]. In order to assess the construct validity of Questionnaire A, we followed the approach employed in Adelman [1] and we compared the results of Questionnaire A to the responses to Questionnaire B. Questionnaire B was open-ended and the subjects had to answer what they liked or disliked most about the ESY. When we mapped the answers to Questionnaire B to concepts or evaluation criteria in Questionnaire A, we found that the subjects disliked aspects of the ESY related to evaluation criteria with low scores and they liked aspects referring to criteria with high scores (see Table 2). For example, some subjects did not understand the sensitivity analysis graph (understanding) and they found the user interface somehow complex (ease of use). Many subjects, however, acknowledged that the ESY helps the DMs perform decision-making tasks effectively (performance) and that it provides a broad variety of functionalities and outputs (completeness). These results indicate that the two questionnaires are related.

Appendix D. Questionnaire B (open-ended)

Please read carefully and answer the following questions:

1. What did you like and/or find most useful about the ESY? Why?
2. What did you dislike and/or find most restrictive or ineffective about the ESY? Why?
3. What specific changes and/or modifications would you suggest regarding the following features of the ESY? Please write NONE if you do not have any suggestions for improving the particular features.
   - 3.1. The knowledge representation (e.g. constraints):
   - 3.2. The generation of strategies (Coarse expert system):
   - 3.3. The ranking of strategies (Ranking module):
   - 3.4. The explanation mechanism (Fine expert system):
   - 3.5. The sensitivity analysis tools:
   - 3.6. User interface:
   - 3.7. Graphic displays:
   - 3.8. Other:

4. What do you think would be the future potential of the ESY in the event of a radiation accident? Why?
5. What do you think would be the future potential of using the ESY for training purposes? Why?
6. Where else do you think an operational version of the ESY would receive good acceptance? Why?
7. Please give any other comments that you feel are relevant to this questionnaire.

References

[1] L. Adelman, Evaluating Decision Support and Expert Systems, Wiley, New York, 1992.

[2] J.E. Bailey, S.W. Pearson, Development of a tool for measuring and analyzing computer user satisfaction, Management Science 29 (1983) 530–545.

[3] J.F. Bard, T.A. Feo, S.D. Holland, Reengineering and the development of a decision support system for printed wiring board assembly, IEEE Transactions on Engineering Management 42 (1995) 91–98.

[4] J. Bartzis, J. Ehrhardt, S. French, J. Lochard, M. Morrey, K.N. Papamichail, K. Sinkko, A. Sohier, RODOS: decision support for nuclear emergencies, in: S.H. Zanakis, G. Doukidis, C. Zopounidis (Eds.), Recent Developments and Applications in Decision Making, Kluwer Academic Publishers, Dordrecht, The Netherlands, 2000, pp. 379–394.

[5] S. Belardo, K.R. Karwan, W. Wallace, An investigation of system-design considerations for emergency management decision support, IEEE Transactions on Systems, Man, and Cybernetics 14 (1984) 795–804.

[6] A. Blair, J. Debenham, J. Edwards, A comparative study of methodologies for designing IDSS, European Journal of Operational Research 103 (1997) 277–295.

[7] D. Borenstein, Towards a practical method to validate Decision Support Systems, Decision Support Systems 23 (1998) 227–239.

[8] J.E. Boritz, A.K.P. Wensley, Evaluating expert systems with complex outputs—the case of audit planning, Auditing-a Journal of Practice and Theory 11 (1992) 14–29.

[9] E.G. Carmines, R.A. Zeller, Reliability and Validity Assessment, Sage Publications, Beverly Hill, CA, 1979.

[10] D.H.J. Caro, Expert support systems in an emergency management environment—an evaluative framework, Educational & Training Technology International 29 (1992) 132–142.

[11] C. Changchit, C.W. Holsapple, R.E. Viator, Transferring auditors' internal control evaluation knowledge to management, Expert Systems with Applications 20 (2001) 275–291.

[12] M. Chau, D. Zeng, H.C. Chen, M. Huang, D. Hendriawan, Design and evaluation of a multi-agent collaborative Web mining system, Decision Support Systems 35 (2003) 167–183.

[13] J.Q. Chen, S.M. Lee, An exploratory cognitive DSS for strategic decision making, Decision Support Systems 36 (2003) 147–160.

[14] A. Clarke, B. Soufi, L. Vassie, J. Tyrer, Field evaluation of a prototype laser safety decision support system, Interacting with Computers 7 (1995) 361–382.

[15] D.C. Cliburn, J.J. Feddema, J.R. Miller, T.A. Slocum, Design and evaluation of a decision support system in a water balance application, Computers & Graphics 26 (2002) 931–949.

[16] J.P. Costa, P. Melo, P. Godinho, L.C. Dias, The AGAP system: a GDSS for project analysis and evaluation, European Journal of Operational Research 145 (2003) 287–303.

[17] L.G. Cronbach, Coefficient alpha and the internal structure of tests, Psychometrica 16 (1951) 297–334.

[18] L.G. Cronbach, Test validation, in: R.L. Thorndike (Ed.), Educational Measurement, American Council of Education, Washington, DC, 1971, pp. 443–507.

[19] F.D. Davis, Perceived usefulness, perceived ease of use, and user acceptance of information technology, MIS Quarterly 13 (1989) 319–340.

[20] R.A. deMillo, W.M. McCracken, R.J. Martin, J.F. Passafiume, Software Testing and Evaluation, Benjamin/Cummings, Menlo Park, 1987.

[21] J.S. Dhaliwal, I. Benbasat, The use and effects of knowledge-based system explanations: theoretical foundations and a framework for empirical evaluation, Information Systems Research 7 (1996) 342–362.

[22] W. Edwards, How to use multiattribute utility measurement for social decisionmaking, IEEE Transactions on Systems, Man, and Cybernetics SMC-7 (5) (1977) 326–340.

[23] J. Ehrhardt, J. Brown, S. French, G.N. Kelly, T. Mikkelsen, H. Muller, RODOS: decision-making support for off-site emergency management after nuclear accidents, Kerntechnik 62 (1997) 122–128.

[24] P.N. Finlay, M. Forghani, A classification of success factors for Decision Support Systems, Journal of Strategic Information Systems 7 (1998) 53–70.

[25] P.N. Finlay, J.M. Wilson, A survey of contingency factors affecting the validation of end-user spreadsheet-based Decision Support Systems, Journal of the Operational Research Society 51 (2000) 949–958.

[26] P.N. Finlay, G.J. Forsey, J.M. Wilson, The validation of expert systems-contrasts with traditional methods, Journal of the Operational Research Society 39 (1988) 933–938.

[27] G.A. Forgionne, An AHP model of DSS effectiveness, European Journal of Information Systems 8 (1999) 95–106.

[28] S.I. Gass, Documenting a computer-based model, Interfaces 14 (1984) 84–93.

[29] S.I. Gass, Model accreditation—a rationale and process for determining a numerical rating, European Journal of Operational Research 66 (1993) 250–258.

[30] L. Geneste, B. Grabot, A. Letouzey, Scheduling uncertain orders in the customer-subcontractor context, European Journal of Operational Research 147 (2003) 297–311.

[31] M. Grabowski, S.D. Sanborn, Evaluation of embedded intelligent real-time systems, Decision Sciences 32 (2001) 95–123.

[32] S. Gregor, I. Benbasat, Explanations from intelligent systems: theoretical foundations and implications for practice, MIS Quarterly 23 (1999) 497–530.

[33] S.A. Guerlain, P.J. Smith, J.H. Obradovich, S. Rudmann, P. Strohm, J.W. Smith, J. Svirbely, L. Sachs, Interactive critiquing as a form of decision support: an empirical evaluation, Human Factors 41 (1999) 72–89.
[34] B. Guijarro-Berdiñas, A. Alonso-Betanzos, Empirical evaluation of a hybrid intelligent monitoring system using different measures of effectiveness, Artificial Intelligence in Medicine 24 (2002) 71–96.
[35] R.P. Hämäläinen, K. Sinkko, M. Lindstedt, M. Amman, A. Salo, RODOS and Decision Conferencing on Early Phase Protective Actions in Finland, STUK Radiation and Nuclear Safety Authority, Helsinki, Finland, 1998.
[36] E. Hatch, H. Farhady, Research Design and Statistics for Applied Linguistics, Newbury House, Rowley, MA, 1982.
[37] P. Herabat, P. Songchitruksa, A decision support system for flexible pavement treatment selection, Computer-Aided Civil and Infrastructure Engineering 18 (2003) 147–160.
[38] M.E. Hernando, E.J. Gomez, R. Corocy, F. del Pozo, Evaluation of DIABNET, a decision support system for therapy planning in gestational diabetes, Computer Methods and Programs in Biomedicine 62 (2000) 235–248.
[39] S. Holtzman, Intelligent Decision Systems, Addison-Wesley, Reading, MA, 1989.
[40] P.J.H. Hu, P.C. Ma, P.Y.K. Chau, Evaluation of user interface designs for information retrieval systems: a computer-based experiment, Decision Support Systems 27 (1999) 125–143.
[41] G.S. Hubona, J.E. Blanton, Evaluating system design features, International Journal of Human-Computer Studies 44 (1996) 93–118.
[42] ICRP, Optimisation and decision-making in radiological protection, Annals of the ICRP, ICRP Publication, vol. 55, 1989, p. 20.
[43] D.R. Insua, E. Gallego, A. Mateos, S. Rios-Insua, MOIRA: a decision support system for decision making on aquatic ecosystems contaminated by radioactive fallout, Annals of Operation Research 95 (2000) 341–364.
[44] S. Kanungo, S. Sharma, P.K. Jain, Evaluation of a decision support system for credit management decisions, Decision Support Systems 30 (2001) 419–436.
[45] R. Kathuria, M. Anandarajan, M. Igbaria, Linking IT applications with manufacturing strategy: an intelligent decision support system approach, Decision Sciences 30 (1999) 959–991.
[46] R.L. Keeney, Siting Energy Facilities, Academic Press, New York, 1980.
[47] R.L. Keeney, Decision-analysis—an overview, Operations Research 30 (1982) 803–838.
[48] R.L. Keeney, H. Raiffa, Decisions with Multiple Objectives: Preferences and Value Tradeoffs, Wiley, New York, 1976.
[49] D.A. Klein, Decision-Analytic Intelligent Systems: Automated Explanation and Knowledge Acquisition, Lawrence Erlbaum Associates, New Jersey, 1994.
[50] K.A.H. Kobbacy, N.C. Proudlove, M.A. Harper, Towards an intelligent maintenance optimization system, Journal of the Operational Research Society 46 (1995) 831–853.
[51] C.T. Kydd, Cognitive biases in the use of computer-based Decision Support Systems, Omega-International Journal of Management Science 17 (1989) 335–344.
[52] S.L. Li, The development of a hybrid intelligent system for developing marketing strategy, Decision Support Systems 27 (2000) 395–409.
[53] S. Liao, Case-based decision support system: architecture for simulating military command and control, European Journal of Operational Research 123 (2000) 558–567.
[54] H.P. Lu, A framework for GSS evaluation: an organizational perspective, International Journal of Technology Management 12 (1996) 221–230.
[55] J.M. Mackay, S.H. Barr, M.G. Kletke, An empirical-investigation of the effects of decision aids on problem-solving processes, Decision Sciences 23 (1992) 648–672.
[56] J.Y. Mao, I. Benbasat, The use of explanations in knowledge-based systems: cognitive perspectives and a process-tracing analysis, Journal of Management Information Systems 17 (2000) 153–179.
[57] M. Martinsons, R. Davison, D. Tse, The balanced scorecard: a foundation for the strategic management of information systems, Decision Support Systems 25 (1999) 71–88.
[58] J. McGovern, D. Samson, A. Wirth, Using case-based reasoning for basis development in intelligent decision systems, European Journal of Operational Research 77 (1994) 40–59.
[59] H.J. Miser, E.S. Quade, Handbook of Systems Analysis—Craft Issues and Procedural Choices, Wiley, USA, 1988.
[60] E. Mosqueira-Rey, V. Moret-Bonillo, Validation of intelligent systems: a critical study and a tool, Expert Systems with Applications 18 (2000) 1–16.
[61] S. Murrell, R.T. Plant, A survey of tools for the validation and verification of knowledge-based systems: 1985–1995, Decision Support Systems 21 (1997) 307–323.
[62] H.R. Nemati, D.M. Steiger, L.S. Iyer, R.T. Herschel, Knowledge warehouse: an architectural integration of knowledge management, decision support, artificial intelligence and data warehousing, Decision Support Systems 33 (2002) 143–161.
[63] M.E. Nissen, Knowledge-based knowledge management in the reengineering domain, Decision Support Systems 27 (1999) 47–65.
[64] M. Norušis, SPSS Professional Statistics 6.1, SPSS, Chicago, 1994.
[65] J.C. Nunnaly, Psychometric Theory, McGraw-Hill, New York, 1967.
[66] R.M. O'Keefe, A.D. Preece, The development, validation and implementation of knowledge-based systems, European Journal of Operational Research 92 (1996) 458–473.
[67] R.M. O'Keefe, O. Balci, E.P. Smith, Validating expert system performance, IEEE Intelligent Systems & Their Applications 2 (1987) 81–90.
[68] D.E. O'Leary, Measuring the quality of computer-model performance, European Journal of Operational Research 56 (1992) 319–331.
[69] K.N. Papamichail, The design of an evaluation system in RODOS and its assessment, in: D.K. Despotis, C. Zopouni-

dis (Eds.), 5th International Conference of the Decision Sciences Institute, New Technologies Publications, Athens, 1999, pp. 439–442.
[70] K.N. Papamichail, Intelligent Decision Support for Nuclear Emergencies, PhD thesis, Department of Computer Science, University of Manchester, Manchester, 2000.
[71] K.N. Papamichail, S. French, Generating feasible strategies in nuclear emergencies—a constraint satisfaction problem, Journal of the Operational Research Society 50 (1999) 617–626.
[72] K.N. Papamichail, S. French, Decision support in nuclear emergencies, Journal of Hazardous Materials 71 (2000) 321–342.
[73] K.N. Papamichail, S. French, Explaining and justifying the advice of a decision support system: a natural language generation approach, Expert Systems with Applications 24 (2003) 35–48.
[74] M. Parikh, B. Fazlollahi, S. Verma, The effectiveness of decisional guidance: an empirical evaluation, Decision Sciences 32 (2001) 303–331.
[75] S.C. Park, S. Piramuthu, M.J. Shaw, Dynamic rule refinement in knowledge-based data mining systems, Decision Support Systems 31 (2001) 205–222.
[76] P. Perny, D. Vanderpooten, An interactive multiobjective procedure for selecting medium-term countermeasures after nuclear accidents, Journal of Multi-Criteria Decision Analysis 7 (1998) 48–60.
[77] L.D. Phillips, Requisite decision modeling—a case-study, Journal of the Operational Research Society 33 (1982) 303–311.
[78] S. Ram, S. Ram, Validation of expert systems for innovation management: issues, methodology, and empirical assessment, Journal of Product Innovation Management 13 (1996) 53–68.
[79] D.C. Ranyard, Decision Support for Nuclear Emergency Response, PhD thesis, School of Computer Studies, University of Leeds, Leeds, 1996.
[80] P.J. Regan, S. Holtzman, Research-and-development decision adviser—an interactive approach to normative decision system model construction, European Journal of Operational Research 84 (1995) 116–133.
[81] E. Reiter, R. Dale, Building Applied Natural Language Generation Systems, Cambridge University Press, Cambridge, UK, 2000.
[82] J.V. Richardson, Question master: an evaluation of a web-based decision-support system for use in reference environments, College & Research Libraries 59 (1998) 29–37.
[83] M. Rungtusanatham, Let's not overlook content validity, Decision Line 29 (1998) 10–13.
[84] V.V.S. Sarma, Decision-making in complex-systems, Systems Practice 7 (1994) 399–407.
[85] A.G.M. Schenker-Wicki, The Conceptual Definition of a Crisis Management Decision Support System CMDSS for Evaluating Acceptable Countermeasures to Reduce Ingestion Dose after an Accidental Release of Radioactivity, University Press Fribourg Switzerland, Fribourg, 1990.
[86] R. Sharda, S.H. Barr, J.C. Mcdonnell, Decision Support System effectiveness—a review and an empirical test, Management Science 34 (1988) 139–159.
[87] R.S. Sharma, D.W. Conrath, D.M. Dilts, A sociotechnical model for deploying expert systems: Part I. The general-theory, IEEE Transactions on Engineering Management 38 (1991) 14–23.
[88] B.G. Silverman, Evaluating and refining expert critiquing systems—a methodology, Decision Sciences 23 (1992) 86–110.
[89] B.G. Silverman, Unifying expert-systems and the Decision Sciences, Operations Research 42 (1994) 393–413.
[90] B.G. Silverman, Knowledge-based systems and the Decision Sciences, Interfaces 25 (1995) 67–82.
[91] R.L. Teach, E.H. Shortliffe, An analysis of physician attitudes regarding computer-based clinical consultation systems, Computers and Biomedical Research 14 (1981) 542–558.
[92] P. Todd, I. Benbasat, An experimental investigation of the relationship between decision-makers, decision aids and decision-making effort, Information Systems Research 31 (1993) 80–100.
[93] E. Tsang, Foundations of Constraint Satisfaction, Academic Press, London, 1993.
[94] L.L. Tung, M.A. Quaddus, Cultural differences explaining the differences in results in GSS: implications for the next decade, Decision Support Systems 33 (2002) 177–199.
[95] E. Turban, J.E. Aronson, Decision Support Systems and Intelligent Systems, Prentice Hall, New Jersey, 2001.
[96] A.M. Turing, Computing machinery and intelligence, Mind 59 (1950) 433–460.
[97] P. Volkner, B. Werners, A decision support system for business process planning, European Journal of Operational Research 125 (2000) 633–647.
[98] C.H. Wu, SODPM: a sequence-oriented decision process model for unstructured group decision problems, Behaviour & Information Technology 21 (2002) 59–69.
[99] J. Zeleznikow, J.R. Nolan, Using soft computing to build real world intelligent Decision Support Systems in uncertain domains, Decision Support Systems 31 (2001) 263–285.

Konstantinia Nadia Papamichail is Lecturer in Information Systems at Manchester Business School. She has previously held posts at the Universities of Leeds and Manchester (UK). Her publications appear in books and journals including Journal of the Operational Research Society, Expert Systems with Applications, Artificial Intelligence, Journal of Hazardous Materials, Omega, Journal of Multi-Criteria Decision Analysis and Lecture Notes in Computer Science. Nadia's work has focused on the design, development and evaluation of Decision Support Systems. Of particular interest is the use of business process modelling for analysing distributed decision processes and improving decision-making practices. Other research interests include intelligence decision support systems, knowledge management systems, environmental decision making and e-learning.

Simon French is Professor of Informatics at Manchester Business School. He was previously Professor of Information Systems and Operational Research at the University of Leeds. With over 100 publications to his name, he has an international reputation in decision theory, analysis and support systems, risk assessment and Bayesian statistics, and is known for his applied work within the nuclear sector, radiation protection and, more generally, societal risk management. For several years, he has played a leading role in a large European project, RODOS, to build a decision support system for nuclear emergency management. Recently, he has worked with NII on the decommissioning of the UK's Magnox reactors. Over the past few years, much of his work has concerned on risk communication per se. He has undertaken research for the UK MAFF (later DEFRA), DH, Cabinet Office and the Food Standard Agency on aspects of the public communication of risk.