Combining Quantitative and Qualitative Measures of Uncertainty in Model-Based Environmental Assessment: The NUSAP System

Jeroen P. van der Sluijs,[1]* Matthieu Craye,[2] Silvio Funtowicz,[2] Penny Kloprogge,[1] Jerry Ravetz,[3] and James Risbey[4]

This article discusses recent experiences with the Numeral Unit Spread Assessment Pedigree (NUSAP) system for multidimensional uncertainty assessment, based on four case studies that vary in complexity. We show that the NUSAP method is applicable not only to relatively simple calculation schemes but also to complex models in a meaningful way and that NUSAP is useful to assess not only parameter uncertainty but also (model) assumptions. A diagnostic diagram can be used to synthesize results of quantitative analysis of parameter sensitivity and qualitative review (pedigree analysis) of parameter strength. It provides an analytic tool to prioritize uncertainties according to quantitative and qualitative insights in the limitations of available knowledge. We show that extension of the pedigree scheme to include societal dimensions of uncertainty, such as problem framing and value-laden assumptions, further promotes reflexivity and collective learning. When used in a deliberative setting, NUSAP pedigree assessment has the potential to foster a deeper social debate and a negotiated management of complex environmental problems.

KEY WORDS: Controversy; NUSAP; pedigree; problem frames; reflexivity; uncertainty; value-laden assumptions

1. INTRODUCTION

Model-based assessment and foresight of complex environmental problems is limited by many different types of uncertainty. The available knowledge base consists of a mixture of (partial) knowledge, assumptions, and ignorance. Policy decisions need to be made before conclusive scientific evidence on these problems is available, while at the same time the potential error costs of wrong decisions can be huge. Usually, controversies surround these problems, in which three interrelated factors play a key role: uncertainty in the knowledge base, differences in framing of the problem, and the inadequacy of the institutional arrangement at the science-policy interface.[(1)] This societal context implies an urgent need for a deliberative, reflexive, and multidimensional approach to uncertainty assessment.[(1–4)] In such an approach the discussion of uncertainty should not be limited to scientists and should take place within a process, taking into account the different perspectives on the problem. Problem framing is seen as a crucial element in uncertainty assessment.

Mainstream uncertainty methods such as Monte Carlo analysis, subjective probability, or Bayesian updating alone are not suitable for this class of problems

[1] Copernicus Institute for Sustainable Development and Innovation, Utrecht University, The Netherlands.

[2] Knowledge Assessment Methodologies (KAM), Institute for the Protection and Security of the Citizen (IPSC), European Commission—Joint Research Centre (EC-JRC), Ispra, Italy.

[3] Research Method Consultancy (RMC), London.

[4] School of Mathematical Sciences, Monash University, Clayton, Australia.

[*] Address correspondence to Jeroen P. van der Sluijs, Copernicus Institute, Utrecht University, Heidelberglaan 2, 3584 CS Utrecht, The Netherlands; j.p.vandersluijs@chem.uu.nl.

0272-4332/05/0100-0481$22.00/1 © 2005 Society for Risk Analysis

because the main problem characteristic is that unquantifiable uncertainties dominate the quantifiable ones. Unquantifiable uncertainties include those associated with problem framings, model structures, assumptions, system boundaries, indeterminacies, and value ladenness. Although quantitative techniques are essential in any uncertainty analysis, they can only account for what can be quantified in a credible way, and thus provide only a partial insight in what usually is a very complex mass of uncertainties. Key dimensions of uncertainty in the knowledge base of complex environmental problems that need to be addressed are technical (inexactness), methodological (unreliability), epistemological (ignorance), and societal (social robustness). Quantitative methods address the technical dimension only. They can, however, be complemented with new qualitative approaches addressing aspects of uncertainty that are hard to quantify and were therefore largely underaddressed in the past. In a number of projects, we have implemented, demonstrated, and tested a novel approach to uncertainty assessment known as the NUSAP method (Numeral Unit Spread Assessment Pedigree) that complements state-of-the-art quantitative uncertainty methods[5,6] with systematic qualitative assessment. This article presents and discusses some of our experiences with the application of the NUSAP method, building on four case studies that vary in complexity. In the first two cases (emission monitoring and emission scenarios), NUSAP is merely used as an analytical device assessing technical, methodological, and epistemic dimensions of uncertainty. In the other two cases (assumptions in quantitative environmental foresight and controversies on environmental health risks), the approach is further extended to cover societal dimensions, such as controversy, problem framing, institutional dimensions, and stakeholder views, in a deliberative and reflexive way.

2. NUSAP AND THE DIAGNOSTIC DIAGRAM

NUSAP is a notational system proposed by Funtowicz and Ravetz,[7] which aims to provide an analysis and diagnosis of uncertainty in the knowledge base of complex (environmental) policy problems. It captures both quantitative and qualitative dimensions of uncertainty and enables one to communicate these in a standardized and self-explanatory way. The basic idea is to qualify quantities using the five qualifiers of the NUSAP acronym: Numeral, Unit, Spread, Assessment, and Pedigree.

We will discuss the five qualifiers. The first is *Numeral*; this will usually be an ordinary number, but when appropriate it can be a more general quantity, such as the expression "a million" (which is not the same as the number lying between 999,999 and 1,000,001). Second comes *Unit*, which may be of the conventional sort, but may also contain extra information, as the date at which the unit is evaluated (most commonly with money). The middle category is *Spread*, which generalizes from the "random error" of experiments or the "variance" of statistics. Although *Spread* is usually conveyed by a number (either ±, %, or "factor of"), it is not an ordinary quantity, for its own inexactness is not of the same sort as that of measurements. Methods to address *Spread* can be statistical data analysis, sensitivity analysis, or Monte Carlo analysis, possibly in combination with expert elicitation.

The remaining two qualifiers constitute the more qualitative side of the NUSAP expression. *Assessment* expresses qualitative judgments about the information. In the case of statistical tests, this might be the significance level; in the case of numerical estimates for policy purposes, it might be the qualifier "optimistic" or "pessimistic." In some experimental fields, information is given with two ± terms, of which the first is the spread, or random error, and the second is the "systematic error," which must estimated on the basis of the history of the measurement, and that corresponds to our assessment. It might be thought that the "systematic error" must always be less than the "experimental error," or else the stated "error bar" would be meaningless or misleading. But the "systematic error" can be well estimated only in retrospect, and then it can give surprises.

Finally, there is P for *Pedigree*, which conveys an evaluative account of the production process of information, and indicates different aspects of the underpinning of the numbers and scientific status of the knowledge used. *Pedigree* is expressed by means of a set of pedigree criteria to assess these different aspects. Assessment of pedigree involves qualitative expert judgment. To minimize arbitrariness and subjectivity in measuring strength, a pedigree matrix is used to code qualitative expert judgments for each criterion into a discrete numeral scale from 0 (weak) to 4 (strong) with linguistic descriptions (modes) of each level on the scale. Each special sort of information has its own aspects that are key to its pedigree; so different pedigree matrices using different pedigree criteria can be used to qualify different sorts of information, as we will see in the case studies below.

In the first two case studies summarized in this article, NUSAP complements quantitative analysis with expert judgment of reliability (Assessment) and systematic multicriteria evaluation of the different phases of production of a given knowledge base (Pedigree). In the other two case studies, the pedigree assessment has been further extended to also address societal dimensions.

NUSAP provides insight on two independent properties related to uncertainty in numbers, namely, spread and strength. *Spread* expresses inexactness whereas *strength* expresses the methodological and epistemological limitations of the underlying knowledge base. The two metrics can be combined in a diagnostic diagram mapping *strength* of, for instance, model parameters and sensitivity of model outcome to *spread* in these model parameters. The diagnostic diagram is based on the notion that neither spread alone nor strength alone is a sufficient measure for quality. Robustness of model output to parameter strength could be good even if parameter strength is low, provided that the model outcome is not critically influenced by the spread in that parameter. In this situation, our ignorance of the true value of the parameter has no immediate consequences because it has a negligible effect on model outputs. Alternatively, model outputs can be robust against parameter spread even if its relative contribution to the total spread in the model is high, provided that parameter strength is also high. In the latter case, the uncertainty in the model outcome adequately reflects the inherent irreducible uncertainty in the system represented by the model. Uncertainty then is a property of the modeled system and does not stem from imperfect knowledge of that system. Mapping components of the knowledge base in a diagnostic diagram thus reveals the weakest spots and helps in the setting of priorities for improvement.

3. EXPERIENCES IN APPLYING THE NUSAP SYSTEM

3.1. Case I: NO_x, SO_2, and NH_3 Emissions in the Netherlands

Emissions of acidifying gases (NO_x, SO_2, and NH_3) in the Netherlands are monitored in the framework of policies on acidification and trans-boundary air pollution. NO_x and SO_2 are mainly the product of combustion of fuels and NH_3 is mainly the product of manure in agriculture. The emission inventory for these gases is based on the detailed monitoring of sources that lead to emissions of these gases. Each source in the inventory is differentiated per activity rate (fuel use or activity data). The inventory distinguishes 419 source-activity combinations (e.g., "highway kilometers of personal cars on gasoline" or "application of manure of dairy cows"). By multiplying each of these detailed activity data with specific emission factors for each activity and each gas, emissions of each gas per source are obtained. These calculated emissions for each source are then added to obtain total annual national emissions of each gas. The emissions for NO_x, SO_2, and NH_3 are then integrated to so-called acidification equivalents (AE) to account for the different contribution that each gas has to acidification.

A comprehensive NUSAP-based uncertainty assessment of the above sketched emission monitoring has been carried out by Gijlswijk *et al.*,[8] which we will briefly summarize here. The analysis followed the following steps: (1) key sources analysis; (2) quantification of probability density functions (PDFs) and pedigree scoring for key sources by expert elicitation; (3) Monte Carlo analysis; (4) combination of Monte Carlo and pedigree analysis in a diagnostic diagram.

The key source analysis allowed a ranking of the source-activity combinations according to contribution to emission total in 2000 and trend during 1990–2000. We focused the analysis on those source-activity combinations that were either responsible for 95% of the total AE emission in 2000 or for 95% of the trend in AE emissions from 1990 to 2000. To further simplify and streamline the expert elicitation of PDFs and pedigree, the list was clustered into groups of source-activity combinations with the same common ground. The common ground can, for instance, be the same basic statistical data set or the same emission estimation methodology. The clusters were prioritized using the outcome of the key source analysis and experts were identified for each cluster. The elicitation used the protocol by Risbey *et al.*[9] The pedigree matrix we used is given in Table I.

The elicitation yielded results for 31 clusters, covering together about 160 source-activity combinations including all key sources. For some source-activity combinations, PDFs and pedigree scores were obtained only for activity data or only for emission factors. For source-activity combinations for which no elicitation results were obtained, conservative default estimates were used, which were taken from the Good Practice Guidance for CLRTAP Emission Inventories that provides an uncertainty class per SNAP category (Selected Nomenclature for Air Pollution).[12] About 20 qualitative descriptions of correlations between monitoring input data were identified during the elicitation and implemented in the Monte Carlo calculations.

Table II presents an aggregated summary of pedigree scores for different data types in the monitoring.

Table I. Pedigree Matrix for Emission Monitoring Data[9–11]

Score	Proxy	Empirical	Method	Validation
4	Exact measure	Large sample direct measurements	Best available practice	Compared with independent measurements of same variable
3	Good fit or measure	Small sample direct measurements	Reliable method commonly accepted	Compared with independent measurements of closely related variable
2	Well correlated	Modeled/derived data	Acceptable method limited consensus on reliability	Compared with measurements not independent
1	Weak correlation	Educated guesses/rule of thumb estimate	Preliminary methods unknown reliability	Weak/indirect validation
0	Not clearly related	Crude speculation	No discernible rigor	No validation

It shows that validation scores are poor for all data types. The table further shows that in general the knowledge base for activity data is stronger than the knowledge base for emission factors.

Fig. 1 presents the results of the analysis for AE in a diagnostic diagram. The rank correlations squared that resulted from the Monte Carlo assessment express the sensitivity of total emission to inexactness in input data, whereas strength (measured by averaged pedigree scores) expresses the methodological and epistemological limitations of the underlying knowledge base.

The diagram clearly identifies three source-activity combinations as the most problematic (i.e., high contribution to overall uncertainty combined with a weak knowledge base, the upper right corner of the diagram).

3.2. Case II: A Complex Model

The Targets IMage Energy Regional model (TIMER) model is part of the Netherlands Environmental Assessment Agency's (RIVM) Integrated Model to Assess the Global Environment (IMAGE). TIMER is one of the energy models used for the 2001 greenhouse gas emission scenarios from the Inter Governmental Panel on Climate Change (IPCC). We used the so-called B1 scenario produced with TIMER for the IPCC Special Report on Emissions Scenarios as the case study.

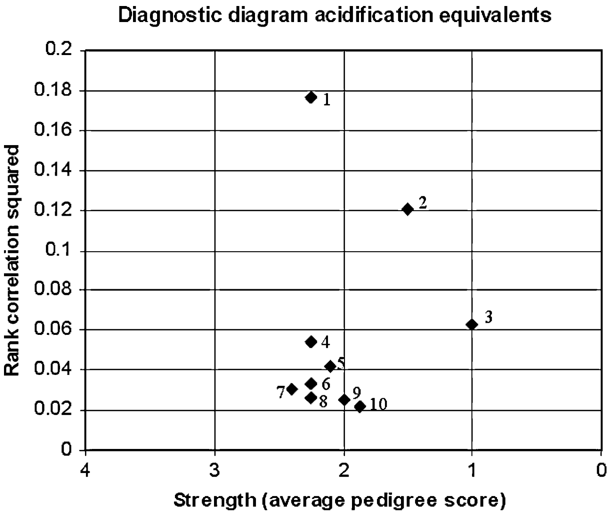
Table II. Average Pedigree Scores for Different Data Types*

	Proxy	Empirical	Method	Validation
Activity data	*2.7 (0.5)*	2.4 (0.7)	*2.6 (0.6)*	**1 (0.9)**
Emission factor NO_x	2.2 (1.1)	2.1 (0.6)	2.5 (0.8)	**1.4 (1.3)**
Emission factor SO_2	*2.6 (1.4)*	2.3 (0.9)	1.7 (0.7)	**1.1 (1.2)**
Emission aggregate NO_x	2.3 (0.6)	*2.6 (0.9)*	2.5 (0.6)	**0.6 (0.7)**
Emission aggregate NH_3	*2.7 (0.7)*	1.4 (0.6)	2.3 (0.5)	2 (0)

*Standard deviation is given in parenthesis.

Scores <1.4 in bold font (poor), 1.4–2.6 in normal font (medium); >2.6 italic font (good).

Using the Morris[13] method for global sensitivity analysis, we explored quantitative uncertainty in parameters in terms of their relative importance in influencing model results. TIMER is a nonlinear model containing a large number of input variables. The Morris method varies parameters one step at a time in such a way that if sensitivity of one parameter is contingent on the values that other parameters may take, the method is likely to capture such dependencies.



Labels of source-activity combinations plotted:

1. NH_3 dairy cows, application of manure
2. NOx mobile sources agriculture
3. NOx agricultural soils
4. NH_3 meat pigs, application of manure
5. NOx highway: gasoline personal cars
6. NH_3 dairy cows, animal housings, and storage
7. NOx highway: truck trailers
8. NH_3 breeding stock pigs, application of manure
9. NH_3 calves, yearlings, application of manure
10. NH_3 application of synthetic fertilizer

Fig. 1. Diagnostic diagram for the 10 most sensitive source-activity combinations for total emission of acidification equivalents.

TIMER contains 300 variables that were all varied over a range from 0.5 to 1.5 times the default values. The method and full results are documented in detail in van der Sluijs *et al.*[14]

The analysis differentiated clearly between sensitive and less sensitive model components. Also, sensitivity to uncertainty in a large number of parameters turned out to be contingent on the particular combinations of samplings for other parameters, reflecting the nonlinear nature of the model. Input variables and model components identified as most sensitive with regard to model output (projected CO_2 emissions) were:

- Population levels and economic activity,
- Variables related to the formulation of intrasectoral structural change of an economy,
- Progress ratios to simulate technological improvements, used throughout the model,
- Variables related to resources of fossil fuels (size and cost supply curves),
- Variables related to autonomous and price-induced energy efficiency improvement, and
- Variables related to initial costs and depletion of renewables.

We assessed parameter pedigree by means of a NUSAP expert elicitation workshop. Nineteen experts in the fields of energy economy and energy systems analysis and uncertainty assessment attended the workshop. We limited the elicitation to those parameters identified either as sensitive by the Morris analysis or as a "key uncertain parameter" in a interview with one of the modelers. Our selection of variables to address in the NUSAP workshop counted 39 parameters. To further simplify the task of reviewing parameter pedigree, we grouped together similar parameters for which pedigree scores might be to some extent similar. This resulted in 18 clusters of parameters. For each cluster a pedigree-scoring card was made, providing definitions and elaborations on the parameters and associated concepts, and a scoring part to fill out the pedigree scores for each parameter. We used the same criteria and pedigree matrix as in the acidifying emissions case (Table I), but added a fifth criterion: *theoretical understanding*. This is because the theoretical understanding of the dynamics of the energy system is in its early stage of development. The modes for this pedigree criterion are: well-established theory (4); accepted theory partial in nature (3); partial theory limited consensus on reliability (2); preliminary theory (1); and crude speculation (0).

For the expert elicitation session, we divided the participants into three parallel groups. Each participant received a set with all 18 cards. Assessment of parameter strength was done by discussing each of the parameters (one card at a time) in a moderated group discussion addressing strengths and weaknesses in the underpinning of each parameter, focusing on, but not restricted to, the five pedigree criteria. In addition, we asked participants to provide a characterization of potential value ladenness. A parameter is said to be potentially value laden when its estimate may well be influenced by one's preferences, perspectives, optimism or pessimism, or co-determined by political or strategic considerations. Participants were asked to draft their pedigree assessment as an *individual* expert judgment, informed by the group discussion.

We used radar diagrams and kite diagrams[9] to graphically represent results (Fig. 2). Both representations use polygons with one axis for each criterion, having 0 in the center of the polygon and 4 on each corner point of the polygon. In the radar diagrams, a line connecting the scores represents the scoring of each expert. The kite diagrams follow a traffic light analogy. The minimum scores in each group for each pedigree criterion span the green kite; the maximum scores span the amber kite. The remaining area is red. The width of the amber band represents expert disagreement on the pedigree scores. In some cases the size of the green area was strongly influenced by a single deviating low score given by one of the experts.

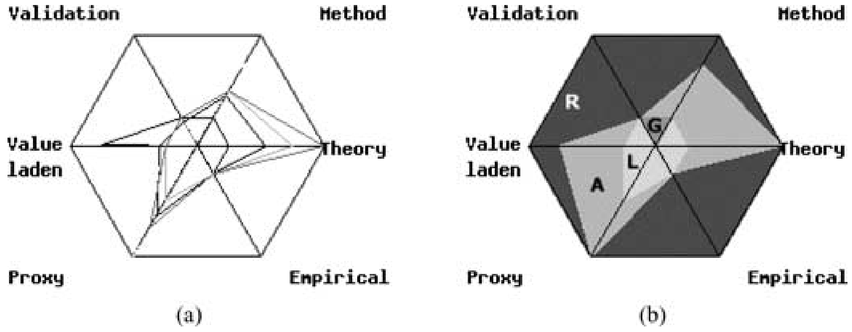


Fig. 2. (a) Example of radar diagram of the gas depletion multiplier assessed by six experts. (b) Same, but represented as kite diagram. G = green, L = light green, A = amber, R = red.

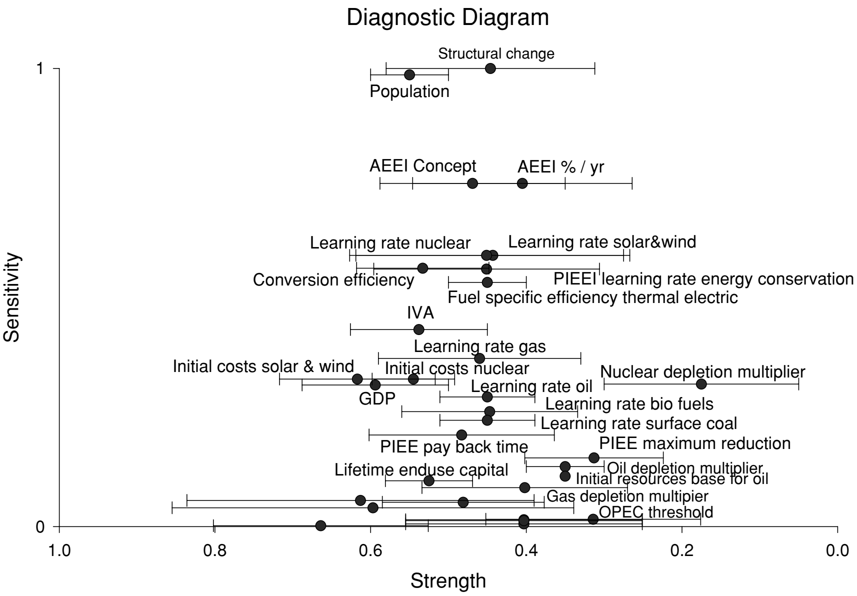


Fig. 3. Diagnostic diagram for key uncertainties in TIMER model parameters.

In those cases the light green kite shows what the green kite would look like if that outlier had been omitted. A kite diagram captures the information from all experts in the group without the need to average expert opinion.

Results from the sensitivity analysis and strength assessments were combined in Fig. 3 to produce a diagnostic diagram.

The diagram shows each of the reviewed parameters plotted. The sensitivity axis measures (normalized) importance of quantitative parameter uncertainty. The strength axis displays the normalized average pedigree scores. Error bars indicate one standard deviation about the average expert value, to reflect expert disagreement on pedigree scores. The strength axis has 1 at the origin and 0 on the right. In this way, the more "dangerous" variables are in the top right quadrant of the plot (high sensitivity, low strength).

We identified three parameters as being close to the danger zone: structural change, B1 population scenario, and autonomous energy efficiency improvement (AEEI). These variables have a large bearing on the CO_2 emission result but have only weak to moderate strength as judged from the pedigree exercise.

Variables that are particularly low in strength also need attention because the theory, data, and method underlying their representation may be weak and we can then expect that they are less perfectly represented in the model. With such high uncertainty in their representation, it cannot be excluded that a better representation would give rise to a higher sensitivity.

This has been the first test of the use of NUSAP on a model of such complexity. The results give enough support to the thought that the method can usefully be adapted and used for other complex model applications as well. An evaluative survey held after the workshop supports this view: the participants appreciated the workshop as a learning experience and unanimously answered the question whether they would like to see this type of NUSAP workshop further applied with "Yes." The overall judgment of the usefulness of the NUSAP workshop by the respondents to the survey was useful (62%) to very useful (38%).

3.3. Case III: Chains of Models

The third case focuses on uncertainty in quantitative environmental indicators based on calculations with a whole chain of softly linked models. As input for the Netherlands Environmental Policy Plan, the Netherlands Environmental Assessment Agency (EAA/RIVM) prepares every four years an assessment of key environmental indicators outlining different future scenarios for a time period of 30 years: the National Environmental Outlook (EO). It presents hundreds of indicators reflecting the pressures on, and future states of, the Dutch, European, and global environments. In a "model chain" of soft-linked computer models—varying in complexity—effects regarding climate, nature and biodiversity, health and safety, and the living environment are calculated for different scenarios. The total of model and other calculations and operations can be seen as a "calculation chain." Often, these chains behind indicators involve many analysts from several departments within the RIVM. Many assumptions have to be made in combining research results in these calculation chains, especially since the output of one computer model often does not fit the requirements of input for the next model (scales, aggregation levels).

We developed a NUSAP-based method to systematically identify, prioritize, and analyze importance and strength of assumptions in these model chains, including potential value ladenness. We demonstrated and tested the method on two Fifth Environmental Outlook (EO5) indicators: "change in length of the growth season" and "deaths and emergency hospital admittances due to tropospheric ozone."

We identified implicit and explicit assumptions in the calculation chain by systematic mapping and deconstruction of the calculation chain, based on document analysis, interviews, and critical review. The resulting list of key assumptions was reviewed and completed in a workshop. Ideally, importance of assumptions should be assessed based on a sensitivity analysis. However, a full sensitivity analysis was not attainable because varying assumptions is much more complicated than, for instance, changing a parameter value over a range: it often requires construction of a new model. Instead, we used the expert elicitation workshop not only to review pedigree of assumptions but also to estimate their quantitative importance.

Table III presents the pedigree matrix used in this study. In the workshop, the experts indicated on scoring cards (one card for each assumption) how they judge the assumption on the pedigree criteria and how much influence they think the assumption has on results. An essential part of our method is that a moderated group discussion takes place in which arguments for high or low scores per criterion are exchanged and discussed. In this way experts in the group remedy each other's blind spots, which enriches the quality of the individual expert judgments. We deliberately did not ask a consensus judgment of the group because we consider expert disagreement a relevant dimension of uncertainty.

Assumptions that have at the same time a high influence on the outcomes of interest and a low pedigree can be qualified as "weak links" in the chain of which the user of the assessment results needs to be particularly aware.

Analysis of the calculation chain of the indicator "deaths and hospital admittances due to exposure to ozone" yielded a list of 24 assumptions. Fourteen key assumptions were selected by the workshop participants as the most important ones, and prioritized. Combining the results of pedigree analysis and estimated influence, the following assumptions showed up as the weakest links of the calculation chain: assumption that uncertainty in the indicator is only determined by the uncertainty in the relative risk (RR is the probability of developing a disease in an exposed group relative to those of a nonexposed group as a function of ozone exposure) and the assumption that the global background concentration of ozone is constant over the 30-year time horizon. The full EO5 case and method for the review of assumptions are documented in Kloprogge *et al.*[15]

3.4. Case IV: Interactive Assessment of Uncertainty in Environmental Health Risk Science and Policy

Near the City of Antwerp, an intense controversy has developed on the potential health effects

Table III. Pedigree Matrix for Reviewing the Knowledge Base of Assumptions

	Score		
Criterion	2	1	0
Plausibility	Plausible	Acceptable	Fictive or speculative
Intersubjectivity peers	Many would make same assumption	Several would make same assumption	Few would make same assumption
Intersubjectivity stakeholders	Many would make same assumption	Several would make same assumption	Few would make same assumption
Choice space	Hardly any alternative assumptions available	Limited choice from alternative assumptions	Ample choice from alternative assumptions
Influence situational limitations (time, money, etc.)	Choice assumption hardly influenced	Choice assumption moderately influenced	Totally different assumption when no limitations
Sensitivity to view and interests of the analyst	Choice assumption hardly sensitive	Choice assumption moderately sensitive	Choice assumption sensitive
Influence on results	Only local influence	Greatly determines the results of link in chain	Greatly determines the results of the indicator

of a waste incinerator. In a neighborhood near the incinerator, an unusually high number of children had congenital defects. Local population and health workers pointed to the incinerator's (dioxin) emissions as the cause. The incinerator's management, supported by local authorities, deemed these accusations as "irrational, meaning purely hypothetical and not scientifically proven."

Following years of heated debate, involving citizen's committees, policymakers (both local and regional), and scientific experts, the conflict evolved to a phase in which all parties realized that a business as usual style will not work any longer. This led to the establishment of the Flemish Centre of Expertise on Environment and Health (CEEH) and initiatives to renew interactions between science, policy, and society.(16)

Against this background, a workshop was held to explore how NUSAP pedigree schemes can support and structure deliberations on uncertainties in environmental risk assessment.(17) The workshop involved experts and actors directly involved and external experts from RIVM and representatives of stakeholders in environmental health issues. The workshop focused in parallel groups on three scientific studies that had been used in the sociopolitical discussions on the incinerator's impact on the environment and local health:

- An epidemiological study that investigated whether there were increased health risks among children whose parents lived or had lived in the particular neighborhood;(18)
- An exposure assessment to estimate the intake of dioxin in the neighborhood around the incinerator during a specified period;(19)
- A biomonitoring study comparing the neighborhood around the incinerator with other industrialized and rural areas, using markers for exposure and effect.(20,21)

Tailored pedigree matrices were designed for each of the studies. The pedigree matrix that was used to structure the deliberative uncertainty assessment in the workshop for the epidemiological study is presented in Table IV

Through analysis of the study reports and interviews with their authors, the main phases in knowledge production to be covered in the pedigree assessment were devised. The choice of these phases reflects the complementarity between more cognitive and more social aspects. A phase related to problem framing was explicitly added. In this way, a discussion was triggered on the "status" of the used problem definition in relation to other disciplinary framings and sociopolitical perspectives and on the "process" through which the expert framing and other sociopolitical framings had (not) been matched.

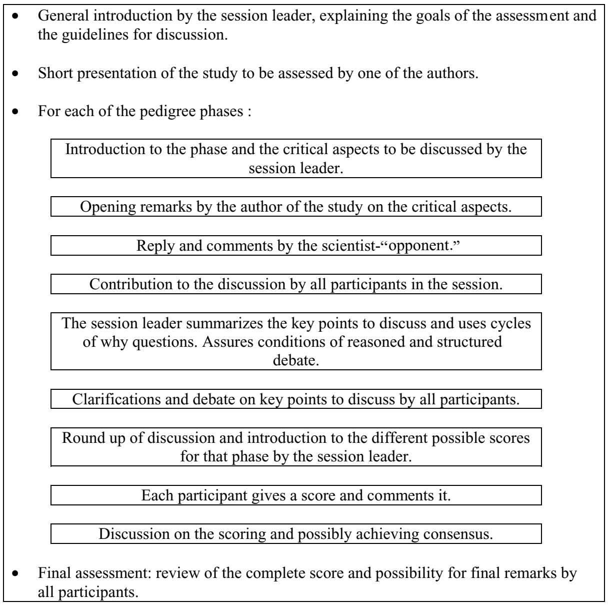
The discussions were shaped as a reasoned, structured debate focusing on underlying assumptions and frame-dependent choices in the different studies. In each session, a discussion leader had to prevent the condition where only technical features of uncertainty would be covered. Fig. 4 synthesizes the protocol followed in the three sessions.

The panel had to assess the "study as a whole" that made it impossible to display on a detailed level all the types of uncertainty involved. Two illustrative critical aspects were presented for each of the phases. These related to choices that had been made in the framing and the design of the study, and subsequently had been criticized by other experts and relevant actors (e.g., under the phase "data-definitions": "how to define the exposed population?"). Also included were other aspects that had not been openly debated in the past but could have led to a more reflexive knowledge development had they been approached with openness (e.g., under the phase "data-definition": "who is competent to define a congenital defect? a family doctor, a parent, a professor in epidemiology, an operator of a database . . . ?").

Two experts—the author of the relevant study and an "opponent" or "critical judge"—introduced each topic. Then, the discussion extended to include the views and reactions of the stakeholders, citizens, and policymakers in the panel. The session leader and

Score	Problem Framing	Data Definitions	Data Collection	Analysis	Review
4	Negotiation	Negotiation	Task force	Established	Extended
3	Scientific	Science	Direct	Discussion	External
2	Compromise	Pragmatic	Bureaucratic	Competition	Independent
1	Inertia	Symbolic	Indirect	Embryonic	Internal
0	Controversy	Unknown	Fiat	No info	None

Table IV. Pedigree Matrix for the Epidemiological Study



- General introduction by the session leader, explaining the goals of the assessment and the guidelines for discussion.
- Short presentation of the study to be assessed by one of the authors.
- For each of the pedigree phases :

Introduction to the phase and the critical aspects to be discussed by the session leader.

Opening remarks by the author of the study on the critical aspects.

Reply and comments by the scientist-"opponent."

Contribution to the discussion by all participants in the session.

The session leader summarizes the key points to discuss and uses cycles of why questions. Assures conditions of reasoned and structured debate.

Clarifications and debate on key points to discuss by all participants.

Round up of discussion and introduction to the different possible scores for that phase by the session leader.

Each participant gives a score and comments it.

Discussion on the scoring and possibly achieving consensus.

- Final assessment: review of the complete score and possibility for final remarks by all participants.

Fig. 4. NUSAP workshop protocol.

another social scientist had to guarantee that an informed and fair debate took place. He also had a list of possible questions in order to (re)focus the discussion if necessary. These model questions were based on insights on the structure of argumentations,[22,23] the content of actors' frames of meaning,[24] and the different types of scientific debate and controversy when uncertainty is salient.[25] In these ways, the protocol enhanced the reflexivity of the process, both in terms of content, that is, opening up the problem definition and the scope of argumentation, and in terms of process, that is, placing the participants in new roles and rules of interaction. This setting challenged the traditional division between the scientist as a provider of facts versus policymakers and the public as defenders of values. The questions promoted discussion on the validity of assumptions, which could reveal particular biases in the framing of the risk. They were intended to deliver insight in the deeper debate on plausible hypotheses, distinguishing it from the more factual discussions on the empirical basis and the methodological work. Included were cycles of typical why questions, for example, "What is the right (research) approach to this problem? . . . Why is this the adequate approach (asks for the definition of the (research) problem)? . . . Why do you define the (research) problem in this way (asks for underlying and supporting "theories")? . . . Why do you use these theories in this case (asks for the fundamental features of framing, the preferences, and convictions)?"

The discussion of each pedigree phase was concluded by giving a score. The scoring was a collective exercise of deliberation that enabled summarizing the main points of discussion, explaining disagreement, and clarifying any ambiguity in the pedigree scheme. The resulting pedigree score for the epidemiological study was (1–2, 1–2, 2, 2–3, 0). The notation n-m signifies disagreement in the group. The low pedigree score of the epidemiological study was consistent with its failure to deliver robust insights and to play a relevant role in the policy debate.

Whereas the problem definition used in the epidemiological study and the choice of data sources and methods of data collection have been intensively discussed between the research team and the Ministry of Public Health, the resulting framing failed to address the concerns of the local population and was quite meaningless from the perspective of the incinerator's management.

The reactions on this framing ranged from "an inadequate use of epidemiology" to "a complete irrelevance of the epidemiological approach." The problem definition used implicitly called into question the existence of the cluster of congenital diseases in the neighborhood by statistically testing the significance of these diseases' incidence in the area compared to the whole Flemish region. Opponents of the study argued a more correct and relevant use of epidemiology would have been to test the relation between these diseases and possible causing factors.

It turned out that during the discussions on the other phases the participants often referred to the frame dependency of certain choices, thus confirming the crucial importance of problem frames. This proved somehow that in the still emerging environmental health science, ignorance and indeterminacy are the predominant forms of uncertainty, largely outweighing in importance methodological and technical aspects.

Overall, the session confirmed the centrality of the issue of framing in this kind of environmental health risk assessment. Participants took more time to discuss the framing than any other phase. The nonscientists also felt that their contribution was most relevant with respect to framing and felt lesser need to intervene in the more "technical" phases dealing with the choices of data sources and of methods. However, they remained very interested and followed with attention the expert discussions on these issues.

The session raised awareness about the complexity of the issues to be studied and the resulting inherent uncertainty and ignorance. As participants learned that choices and assumptions could not be based exclusively on objective science, questions were raised about who is competent and "entitled" to make the necessary choices. In this sense, the session promoted reflexivity and collective learning. It showed the potential of the pedigree assessment to foster a deeper social debate and a negotiated management of environmental health risks.

Many participants suggested the approach could be applied in a constructive way, that is, when policy-supporting research is being developed. However, others argued that the method was still too science-centered, thereby devaluing the contributions by citizens and other lay knowledge providers. The lessons learnt during the workshop are being used to develop a set of pedigree schemes that can be deployed in distinct processes dealing with framing, research design, and extended review. The experience also points to the need to reflect on the integration of these processes and their results in an overall inclusive approach.

4. DISCUSSION

In practice, as we saw most prominently in the TIMER case study, different experts may attribute different pedigree scores to the same part of a given knowledge base. Differences in judgments amongst experts regarding which mode of each column of a pedigree matrix best represents the state of knowledge may stem from different causes. It can be that the experts have different background knowledge on which they base their judgment, it can be that experts interpret the linguistic descriptions in the pedigree matrix differently, and it can be that the experts disagree on a more fundamental level on pedigree scores. The first two causes need to be avoided. They can be minimized by a procedure in which a group discussion between the experts involved in the scoring precedes the scoring, so that information is shared amongst the experts and interpretation issues are discussed so that a shared understanding is achieved. Such a procedure is discussed in detail in van der Sluijs *et al.*[14] The third cause for diverging scores, expert disagreement on pedigree scores, is valuable uncertainty information because it indicates the existence of epistemic uncertainty, such as competing schools of thought within the scientific peer community. Therefore, the reasons for expert disagreement should be explored and information on the disagreement should be preserved in the presentation of results, as we did, for instance, in the kite diagrams (Fig. 2b). One should be very reluctant about averaging pedigree scores elicited from different experts. If the disagreement follows two separate, well-articulated paradigms/arguments, then one would not want to average them. Rather, one would present the pedigree scores separately for each view, noting the reasons given for differences.

A concern put forward by a reviewer of this article is that in this reflexive science approach the concern for process seems to dominate concerns about outcomes of scientific assessments. This raises the question how far one can and should go in deconstructing science, in view of the problems that may occur when different actors in the policy process strategically misuse uncertainty and pedigree to challenge science that does not fit their interests and agendas.

However, the reason why all this procedure is introduced in Europe is that there has been a decline in public trust of science when it is employed in policy processes. Loss of trust in science has been most dramatically manifest in the United Kingdom, particularly after the Bovine Spongiform Encephalopathy (BSE or mad cow disease) crisis. An earlier example is the crisis in the mid 1980s at the International Institute for Applied Systems Analysis (IIASA) when the credibility of its influential energy scenarios was openly challenged. In their critical review of the IIASA energy scenarios, Keepin and Wynne[26] speak of "informal guesswork" and a lack of peer review and quality control, "raising questions about political bias in scientific analysis." More recent, the Netherlands Environmental Assessment Agency (RIVM/MNP) faced a similar crisis after an employee published an article in a newspaper claiming that RIVM's environmental foresight studies are misleading because they are based on the virtual reality of computer models.[4] According to Oreskes *et al.*,[27] we should wonder how much of a model is based "on observation and measurement of accessible phenomena, how much is based on informed judgment, and how much is convenience?"

Unrealistic expectations of science as a provider of certainties increase the potential for loss of trust. NUSAP can promote more realism and a better public understanding of the limits to our capacity to know and understand complex environmental risks.

Our opinion is that, given all the obvious dangers of manipulation of "due process," the dangers of refusing it are even greater. For there could then arise situations where governments' attempt to enforce some scientific policy would lack "the consent of the governed," and there could be substantial loss of trust in institutions.

5. CONCLUSIONS

We have presented experiences and results with the NUSAP method for multidimensional uncertainty assessment in four case studies with increasing complexity: an emission monitoring system, a complex energy model, environmental indicators stemming from calculations with a chain of models, and a major controversy on environmental health risks.

The cases have shown that the NUSAP method is applicable not only to relatively simple calculation schemes but also to complex models in a meaningful way and that it is useful to assess not only parameter uncertainty but also (model) assumptions. A diagnostic diagram synthesizes results of quantitative analysis of parameter sensitivity and qualitative review (pedigree analysis) of parameter strength. It provides a useful analytical tool to prioritize uncertainties according to quantitative and qualitative insights. Extension of the pedigree scheme to include societal dimensions of uncertainty, such as problem framing and value loadings, further promotes reflexivity and collective learning. The task of quality control in the knowledge base of complex and controversial (environmental) policy problems is a complicated one and the NUSAP method disciplines and supports this process by facilitating and structuring a creative reflexive process and in-depth review of the limitations of a given knowledge base. NUSAP promotes to make explicit, and to systematically reflect upon, the various dimensions of uncertainty. It provides a diagnostic tool for assessing the robustness of a given knowledge base for policymaking and promotes criticism by clients and users of all sorts—expert and lay—and will thereby support extended peer-review processes. It helps to focus research efforts on the potentially most problematic parameters and assumptions, identifying at the same time specific weaknesses and biases in the knowledge base.

Similar to a patient information leaflet alerting the patient to risks and unsuitable uses of a medicine, NUSAP enables the delivery of policy-relevant quantitative information together with the essential warnings on its limitations and pitfalls. It thereby promotes the responsible and effective use of the information in policy processes.

REFERENCES

1. Craye, M., Goorden, L., Vandenabeele, J., & Van Gelder, S. (2001). *Milieu en gezondheid in Vlaanderen : naar en adequate dialoog tussen overheid, bevolking en wetenschap*. Antwerp: UFSIA-STEM onderzoeksrapport.
2. Funtowicz, S. O., & Ravetz, J. R. (1993). Science for the post-normal age. *Futures*, *25*(7), 735–755.
3. Van der Sluijs, J. P. (1997). *Anchoring amid uncertainty; On the management of uncertainties in risk assessment of anthropogenic climate change*. Ph.D. thesis, Utrecht University.
4. Van der Sluijs, J. P. (2002). A way out of the credibility crisis of models used in integrated environmental assessment. *Futures*, *34*, 133–146.
5. Saltelli, A., & Tarantola, S. (2002). On the relative importance of input factors in mathematical models: Safety assessment for nuclear waste disposal. *Journal of American Statistical Association*, *97*(459), 702–709.
6. Saltelli, A., Tarantola, S., Campolongo, F., & Ratto, M. (2004). *Sensitivity Analysis in Practice. A Guide to Assessing Scientific Models*. New York: John Wiley.
7. Funtowicz, S. O., & Ravetz, J. R. (1990). *Uncertainty and Quality in Science for Policy*. Dordrecht: Kluwer.
8. Van Gijlswijk, R., Coenen, P., Pulles, T., & van der Sluijs, J. P. (2004). *Uncertainty Assessment of NOx, SO_2 and NH_3 Emissions in the Netherlands*. TNO and Copernicus Institute Research Report. Available at www.nusap.net.
9. Risbey, J. S., Van der Sluijs, J. P., & Ravetz, J. R. (2001). *Protocol for Assessment of Uncertainty and Strength of Emission Data*. Report no. E-2001-10. Dep't of Science Technology and Society, Utrecht University, Utrecht. Available at www.nusap.net.
10. Ellis, E., Gang Li, R., Zhang Yang, L., & Cheng, X. (2000). Long-term change in village-scale ecosystems in China using landscape and statistical methods. *Ecological Applications*, *10*(4), 1057–1089.
11. Van der Sluijs, J. P., Risbey, J. S., & Ravetz, J. R. (in press). Uncertainty assessment of VOC emissions from paint in the Netherlands. *Environmental Monitoring and Assessment*.
12. EEA. (2003). Good practice guidance for CLRTAP emission inventories. In *EMEP/CORINAIR Emission Inventory Guidebook*, 3rd ed. Copenhagen: European Topic Centre on Air and Climate Change (ETC/ACC), European Environment Agency. Available at http://reports.eea.eu.int/EMEPCORINAIR4/en.
13. Morris, M. D. (1991). Factorial sampling plans for preliminary computational experiments. *Technometrics*, *33*(2), 161–174.
14. Van der Sluijs, J. P., Potting, J., Risbey, J., Van Vuuren, D., De Vries, B., Beusen, A., Heuberger, P., Corral Quintana, S., Funtowicz, S., Kloprogge, P., Nuijten, D., Petersen, A., & Ravetz, J. (2002). *Uncertainty Assessment of the IMAGE/TIMER B1 CO_2 Emissions Scenario, Using the NUSAP Method*. Report no. 410 200 104. Dutch National Research Program on Climate Change, Bilthoven. Available at www.nusap.net.
15. Kloprogge, P., van der Sluijs, J. P., & Petersen, A. (2005). *A Method for the Analysis of Assumptions in Assessments Applied to Two Indicators in the Fifth Dutch Environmental Outlook*. Utrecht: Dep't of Science Technology and Society, Utrecht University.
16. Keune, H., Goorden, L., Mertens, R., & Loots, I. (2003). *Communicatie, Interactie en Reflectie over Biomonitoring Nota Opzet Communicatiestrategie, Lokale contacten en Reflectie. Biomonitoringcampagne*. Antwerp: UA-Steunpunt Milieu & Gezondheid, Luik Sociaal en Gezondheidseconomisch Onderzoek.
17. Craye, M., Funtowicz, S., & van der Sluijs, J. P. (2004). A reflexive approach to dealing with uncertainties in environmental health risk science and policy. *International Journal of Risk Assessment and Management*, *5*(2).

18. Aelvoet, W., Nelen, V., Schoeters, G., Vanoverloop, J., Vlietinck, R., & Wallijn, E. (1998). *Risico op gezondheidsschade bij kinderen van de Neerlandwijk te Wilrijk*. Mol: Vito-onderzoeksrapport 1998/TOX/R/030.
19. Schoeters, G., Cornelis, C., Defre, R., & Nouwen, J. (1998). *Studie van gezondheidsaspecten en gezondheidsrisico's ten gevolge van milieuverontreiniging in de Neerlandwijk te Wilrijk*. Mol: Vito-onderzoeksrapport 1998/TOX/R/0097.
20. Koppen, G., Van Loon, H., Vlietinck, B., & Van Larebeke, N. (2001). *Biomonitoringsonderzoek 50-65j vrouwen en 21-40j mannen: evaluatie van de relatie milieu-gezondheid op basis van meting van biomerkers*. Onderzoeksrapport Programma Milieu en Gezondheid in opdracht van het Ministerie van de Vlaamse Gemeenschap.
21. Staessen, J. (2001). *Koepeltekst biomonitoring bij adolescenten: een gevoelige nieuwe methode voor het inschatten van verontreiniging van het leefmilieu en geassocieerde gezondheidsrisico's*. Onderzoeksrapport Programma Milieu en Gezondheid in opdracht van het Ministerie van de Vlaamse Gemeenschap.
22. Toulmin, S. E. (1958). *The Uses of Argument*. Cambridge: Cambridge University Press.
23. Van de Graaf, H., & Hoppe, R. (2000). *Beleid en Politiek*. Bussum: Coutinho.
24. Grin, J., Van de Graaf, H., & Hoppe, R. (1997). *Technology Assessment Through Interaction—A Guide*. Den Haag: Rathenau Instituut.
25. Von Schomberg, R. (1997). *Argumentatie in de context van een wetenschappelijke controverse*. Universiteit Twente WMW-publicatie 27. Delft: Eburon.
26. Keepin, B., & Wynne, B. (1984). Technical analysis of IIASA energy scenarios. *Nature*, *312*, 691–695.
27. Oreskes, N., Shrader-Frechette, K., & Belitz, K. (1994). Verification, validation, and confirmation of numerical models in the earth sciences. *Science*, *263*, 641–646.