Indicators: The Problems of Uncertainty and Quality

S.O. Funtowicz(*) i J.R. Ravetz(*)

Policy-Related Research: The Problem of Uncertainty

No scientific activity is free from uncertainty ; we may say that the key to a science being "matured" is its success in the control of various sorts of uncertainty that affect its results and predictions. These include *inexactness*', as expressed by significant digits, *unreliability*, as expressed in systematic error, and others as well. No amount of sophisticated apparatus and computer power can replace theoretical understanding or the practical skills of controlling it.

When quantitative information is used to provide inputs for the policy process, as in the case of indicators in the social and environmental fields, the scientist's problems of management of uncertainty are more severe in many ways. *First the original data are very rarely as well controlled as in the laboratory*.. Well-structured theories, normally expected to be available in basic or applied science, are conspicuous by their absence in the context of policy-related research. Furthermore such research is *interdisciplinary, involving fields of varying states of maturity and with very different sorts of practice* in their theoretical, experimental and social dimensions. Scientists must use inputs from fields they do not not know intimately; and so they cannot make the same sensitive judgements of quality that they do in their own subject. The result is that quality-control on the research process is diluted; the quality-assurance of results is weaker; and they command less confidence among users.

The problems of uncertainty in policy-related research are further increased in its public dimension. Science is judged by the public (including decision-makers) by its performance in such sensitive areas as hazardous wastes, radioactive fallout, food additives, and reproduction engineering. All these involve much uncertainty, and also inescapable social and ethical aspects. Simplicity and precision in predictions, or even in safe-limits, are not feasible; yet policy-makers tend to expect straightforward information as inputs to their own decision-making process. They want their statistical indicators to provide certainty.

In such circumstances, the maintenance of confidence in science among policy-makers and the general public becomes increasingly difficult. The issue manifests itself at several levels. The simplest is in the representation of uncertainty in quantitative estimates. For example, in risk assessments, the scientific advisor knows that a prediction like a one-in-a-million chance of a serious accident should be hedged about with statements of many sorts of uncertainty, so as to caution any user about the limits of reliability of the numerical assertions. If these are all expressed in prose, the statement becomes tedious and

(*)The Research Methods Consultancy Ltd, Sheffield, Anglaterra.

incomprehensible to the lay user; but if they are omitted or even given in some simple statistical representation, then the same advisor can be accused of conveying a certainty that is not warranted by the facts.

A worse dilemma is encountered whenever scientists give advice on policy-related issues. In addition to low-frequency hazards, these may be of diffused hazards, such as pesticides or food additives; or of possible large-scale future environmental perturbations, such as greenhouse-efect. Such advice is usually supported by the present or expected behaviour of some critical indicator. *Nature* has stressed this dilemma in a recent article, "Half-truths make sense (almost)" (3). This was a comment on a prediction of the consequences of the greenhouse-effect, using the rise in global mean temperature by 2030 A.D. as the indicator. Any definite advice is liable to go wrong; a prediction of danger will appear alarmist if nothing happens in the short run; while reassurance can be condemned if it retrospectively turns out to be incorrect. Thus the credibility of science, based for so long on the supposed certainty of its conclusions, is endangered by *any* sort of scientific advice on such inherently uncertain issues. But suppose then that the scientific advisor prudently refuses to accept ill-founded expert opinions as a basis for quantitative assessments, and declines to provide definite advice when requested by policy-makers. Then science is seen as not performing its public functions, and its legitimacy is threatened. Is there no way through the horns of this dilemma, in which the credibility and the legitimacy of science are both at risk? As *Nature* puts it, "These are among the trials with which policy research centres must contend. Tell the people that there is a muddle, or give them a clear message that they must man the barricades?" The World Resources Institute's solution of the dilemma; the adoption of a computer model is described by *Nature* as a "cop-out".

Thus in policy-related research, the traditional tools for controlling uncertainty need to be enriched. Solving the scientific advisors' dilemma needs something other than more data or bigger, faster computers. Uncertainty is integral to these problems; it cannot be removed by technical means. It must be managed, through clarity on its different aspects. With an appropriate notational system, we can operationalise several sorts of uncertainty, so that quantitative information is qualified in a convenient and comprehensible form, and the quality of the relevant indicators is assured.

NUSAP: The Elicitation of Uncertainty

The presentation of our notational system begins with an elicitation of *uncertainty*. We can start with the simplest sort, usually expressed by error bars and significant digits. Every set of data has a spread; it is an attribute of any quantity, however derived; it may be considered in some contexts as a tolerance, or as a random error in a calculated measurement. It is the kind of uncertainty that relates most directly to the quantity as stated, and is most familiar to students and even the lay public.

A more complex sort of uncertainty relates to the level of confidence to be placed in a quantitative statement. This usually is represented by the confidence limits (at, say, 95% or 99%). In practice, such judgements are quite diverse; thus safety and reliability estimates are given as "conservative by a factor of n". Or in risk analyses, and futures scenarios, estimates are qualified as "optimistic" or "pessimistic". In laboratory practice, the systematic error in physical quantities, as distinct from the random error or spread, is estimated on an historic basis. Thus it provides a kind of assessment to act as a qualifier on the number, or alternatively (if desired) on the spread.

This assessment is one level up from spread, in its sophistication and variety. We may imagine spread as representing *inexactress*, and assessment as expressing *unreliability* (or degree of *reliability* as appropriate). Our knowledge of the behaviour of the data gives us the spread; and our knowledge of its production of intended use, gives us assessment. But there is yet something more. No process, in the field or in the lab, is completely known. Even accepted physical constants may at any time vary in ways that could not have been predicted. This is the realm of our *ignorance*. It includes all the different sorts of gaps in our knowledge, that are not encompassed in the previous two sorts of uncertainty. This ignorance may merely be of what is significant, as when anomalies in experiments are discounted or neglected. Or it may be deeper, as is appreciated retrospectively when great new theoretical advances are made in science.

Can we say anything useful about that of which we are ignorant? It would seen that by the very definition of ignorance, we cannot. But the boundless sea of ignorance has shores; and those we can

stand on, and map. Let us think of a boundary with ignorance as the last sort of uncertainty that we can now effectively control in practical scientific work. To map this boundary, we describe the state-of-the-art of the field of practice in which our quantity is produced. This is done by an evaluative analytical account; this we call the pedigree of the quantity. By means of a matrix it shows the boundary with ignorance by displaying the degree of strength of crucial components of the process. The nature of the boundary, with its crucial components, will depend on the sorts of operations involved. We have analysed two cases: one for the results of research; and the other, for statistical information.

In the research case the phases (or crucial components) are: Theoretical Structures, Data-Input, Peer-Acceptance and Colleague Consensus (Table 1). If we qualify the Theoretical Structure of the production-process of the information as Computational Model (e.g. weather forecasts), we are implicitly stating that we do not have a Theory-based Model (as a hydrodynamical system); and thus we record the absence of an effective theory. Similarly, if Data-Input is not Experimental (as in traditional laboratory sciences), it can be at best Historic/Field Data, as in environmental research. In the latter case, data is inherently less capable of control; and so it is less effective as an input and check for Theoretical Structures. The components on the social side describe the evaluation of the information in the particular context. Peer-Acceptance of a result will be straightforward in a fully matured field where criteria of quality are agreed; a rough aproximation to this is the referee's judgement on a research paper. Colleague Consensus within a research area describes the social strength of the paradigms in which the information is cast; we can see it in the progression of revolutionary theories as (for example) relativity; from embryonic in 1905, through to all but rebels in the 1920's and all but cranks by the 1950's. (Or there is T.H. Huxley's standard progression for new theories, from heresies to superstitions) (4).

Table 1. The Research-Pedigree matrix

Score	*Theoretical Structures*	*Data-Input*	*Peer-Acceptance*	*Colleague Consensus*
4	Established Theory	Experimental Data	Total	All but cranks
3	Theoretically based Model	Historic/ Field Data	High	All but rebels
2	Computational Model	Calculated Data	Medium	Competing Schools
1	Statistical Processing	Educated Guesses	Low	Embryonic Field
0	Definitions	Uneducated Guesses	None	No Opinion

The other pedigree matrix is for statistical information. This is a generic term covering all the data gathered and processed in sequential operations, designed for eventual use in indices and indicators. In this case, the phases are Definitions & Standards, Data Collection & Analysis, Institutional Culture and Review (Table 2). Under Definitions & Standards we have five modes: Negotiation, Science, Convenience, Symbolism and Inertia. The mode Science refers to those cases where the definitions and standards are based on some background body of scientific knowledge. Other considerations, particularly local interests in a policy problem as well as the limits of practical feasibility, can render the strictly scientific approach inappropriate. Thus there is a use for Negotiation among interested parties. Where the science is inadequate for a full specifiation of the definitions and standards, several alternative modes are possible. We speak of Convenience when those actually doing the work adapt any externally imposed definitions and standards to their particular circumstances. Sometimes, however, the scientific basis is nearly, or completely, irrelevant to the interests of those who ultimately control the institutional task. The need for institutional legitimacy of prestige, or other manifestations of values, may come to dominate over all other considerations; the rhetorical value of the indicator is absolute. Here we speak of Symbolism as the mode in this phase. It can be detected by the presence of definitions and standards, which, however technical their form, are known to be widly inappropriate or simply irrelevant to the task. There is a further mode,

Inertia, to characterise the state where noone now even knows or cares why particular definitions and standards came to be used.

Table 2. The Statistical Information-Pedigree matrix

Score	*Definitions & Standards*	*Data-Collection & Analysis*	*Institutional Culture*	*Review*
4	Negotiation	Task-force	Dialogue	External
3	Science	Direct survey	Accommodation	Independent
2	Convenience	Indirect Estimate	Obedience	Regular
1	Symbolism	Educated guess	Evasion	Occasional
0	Inertia	Fiat	No-contact	None
0	Unknown	Unknown	Unknown	Unknown

The second phase, Data Collection & Analysis, refers to the operation in which field data is gathered and then analyzed and reported. Such work is inevitably a sequential operation; analysis comes after collection and will itself usually include several distinct levels. The last two phases of this pedigree refer explicity to the organizational character of the exercise. Institutional Culture characterise relations between different elements in the sequential operation. Review refers to the formalised activities, such as audit or peer-review, for quality-control.

We can now introduce the full notational system designed for the management of uncertainty in quantitative information. We call it NUSAP; the last three letters in the acronym refer to the spread, assessment and pedigree already discussed. The first two refer to numeral and unit. The first category encompasses the arithmetical system; and the second, the base in which it is appropriately expressed.

NUSAP is a *system* because it is not simply a collection of fixed notations. Rather, it is a set of determinate categories, each of which can be filled by particular notations appropriate to the particular context of application. The names of the five categories (or boxes, or places in a string) make up the acronym NUSAP. By means of this place-value representation, each category can be expressed without need for its explicit identification (this is why we refer to it as a "scheme" of notations). For each category, there are many possibilities for conveying particular desired meanings; for example, the representation of a mass in the various systems can be quite neatly expressed. For "five kilograns" we may write $5x10^3$g for a direct measuring in CGS, or perhaps $5:10^3$g for a measuring in"kilo" grams in CGS; or alternatively 5:Kg in MKS, where Kilogram is the unit. We notice in the second case that the unit is split, with a *standard* (grams) operated on by a *multiplier* (10^3). In the third case, using MKS, the unit is *not* split, as Kilogram is fundamental in that system. These cases show how NUSAP representations can convey variants of meanings that my appear not too different in ordinary practice, but which are conceptually quite distinct. Any particular array of such constituents in the five places (as numbers for numeral and CGS for unit) we call a "notation". Given such a notation, any particular case of representation (as $5:10^3$g) will be an "instance" of the notation.

Similar distinctions can be made in energy studies, where (for example) Kilowatts, Megawatts and Gigawatts are not simply a cascade of units connected by 10^3, but refer to physical and accounting operations at very different levels, and have quite different meanings as indices.

To show the use of NUSAP for characterizing uncertainly in information in policy-related research, we take the example critized by *Nature* (3). The chosen indicator (temperature-rise consequent on greenhouse-effect) as described by the World Resources Institute did not exhibit control of its uncertainties, and so (as *Nature* said) could not command confidence. The original statement was of a rise between 1,6 °C and 4,5° C (in average earth temperature over the next 40 years). In NUSAP, this increase is best displayed as,

$$3: \text{°C},2030 : \pm 50\% : \text{Low} : (2,2,3,1))$$

The first three places are derived directly from the quoted quantitative prediction. The range 1,6° to 4,5° is near enough to 3±50%; and it may be said that our mode of representation is more faithful to the scientific meaning of the datum. It thus displays skill in the management of uncertainty, and would therefore tend to mantain confidence in the scientists making the prediction.

How do we justify the pedigree *(2,2,3,1)* (Using the Research-pedigree matrix of Table 1)? We can elicit its rating by analyzing the World Resources Institute's *model of warming commitment.* Are there effective Theory-based Models for atmospheric CO_2 and its temperature effects? According to *Nature,* not quite; severe uncertainties exist on the time-scales both of millions of years and of days. Hence we have at best, Computational Models. What about the data that are injected into the models as inputs? There is some which are better than Educated Guesses but (as yet) not much obtained through instrumental readings, even as Historic/Field Data. Hence we do best to call them Calculated Data, usually resting indirectly on measurements. Moving now to the social aspects of the pedigree, is the state of development of this field, as expressed by Colleague Consensus, better than Embryonic? To have Competing Schools would require the presence of some realistic theoretically-based models; and these are not yet with us. In such a situation, Peer-Acceptance of the result is not such a critical indicator of its quality; here it seems High. *Nature's* criticisms are on the policy implications of the predictions, rather of the quantity itself.

This mapping of the limits of the state-of-the-art exhibits the boundaries with our ignorance. We do not know as much as we could, had there been Theoretically-based Models and Historic/Field Data. In such a case, predictions would have had more strength, more justified urgency, and perhaps also more information on environmental consequences and remedies. As it is, our knowledge is not quite swamped by our ignorance, but it is still too weak and unfocussed for effective decision-making. Our ignorance in the policy aspects of the problem has scarcely been dispelled. Thus the indicator of the World Resources Institute, when properly expressed, tells us more about our ignorance than about the biosphere.

Our assessment rating is based on those considerations; the reliability of the result is judged in the context of its use rather than of its production. Thus we consider the assessment as of practical effectiveness rather than of scientific reproducibility. Here the comments by *Nature* on our Ignorance concerning policy options become crucial. Even if this particular datum were less uncertain, it would still be ineffective. This is why we give a Low rating in assessment. With this improved characterization of the uncertainties of the information, the policy-maker is better equipped for a decision. This might well be in a call for more focussed research, as on practical remedies and the means and time-scales of their implementation. A useful outcome of the problem and its solution could be a set of indicators, based on explicitly "optimistic: or "pessimistic" assumptions, on which more effective monitoring and planning can be based.

In this way we see how relevant evaluations of quality are expressed through the assessment and pedigree categories. These are cast in terms of the characteristic uncertainties of the information, including the border with ignorance; and they can be expressed in a form most appropriate for the policy problem defining the indicator itself.

Indicators: The Elucidation of Quality

In general discussions of indicators, the term *quality* occurs frequently and in many contexts. For example, we hear of *indicators of quality* and of *qualitative indicators,* and even of *quality of indicators.* The similarities of expression, and the overlap of meanings, lead to some confusions, and conceal differences that are important for the proper definition and use of indicators.

Briefly, *indicators of quality* relate to the goodness (or otherwise) of some state of affairs relevant to policy. The *quality might* be of life, the environment, education, research or whatever. It might seem paradoxical to refer then to *quality of indicators;* but this is a judgement of the goodness of its performance of its function, that function being the representation of the goodness of something else. To distinguish between these two levels is essential for competent deployment and criticism of indicators. Finally, *qualitative indicators* invoke *quality* in the sense of non-quantitative. Of course, the boundary between qualitative and quantitative is vague, with ordinal scales and taxonomies lying midway.

Before looking more closely to the problem of quality, let us clarify the distinction between indicators and indices. An index, is, in its broadest sense, a measure of the magnitude of a variable at one point relative to its value at a base point. It is a statistic that may be gathered as a matter of routine, though

it inevitably reflects the dominant conceptions of reality and its representations. An indicator is used to gauge significant trends in some state of affairs. It may be a single selected index, or it may be defined from several indices; it does not exist in isolation from its policy functions. The distinction between indices and indicators is illustrated in the etymology: the index is a pointer (as the index-finger or forefinger), whereas the indicator is the thing that points to some other thing. Many important indicators are called "indices", because they are routinely collected statistics; the distinction is one of function: thus the "Retail Price Index" may be used as an indicator for inflation.

We can provide examples for each of the different meanings of qualiy in relation to indicators. First, consider measures of "the quality of life". Because personal safety and security are so important to people, and also because a society should be seen to be well-organised, the "crime-rate" is foremost among the indicators used to assess this aspect of quality. An increasing crime-rate is an indicator of a social pathology; but its meaning becomes a matter of competing realities. One may be of a decline in the moral standards of private life; the other, a decline in the fairness of social life. Do we need more of police, parental discipline, welfare or jobs? Needless to say, any chosen indicator derives from indices that depend on categories and procedures which are extremely artificial and varied. It follows that any chosen indicator must reflect a particular conception of reality, and it then also confirms and reinforces it in the mind of the public (including politicians and experts alike).

The quality of an indicator, a judgement made on the goodness of its performance of its functions, can be evaluated on technical as well as broadly political grounds. Thus "monetarist" economics, while a subject of strong debate in general, also encountered the difficulty of defining its crucial indicator of "money supply". In the U.K., both M_3 and M_{zero} were tried, but proved erratic and unreliable (5). The index M_3 survives as a statistic, but is no longer regarded as an indicator.

The problem of the measurement of quality is well illustrated in science policy. Traditionally, scientific achievement was assessed mainly on numbers of publications. But this became too easily abused, in an age of proliferation of journals. Then ever more refined criteria and procedures were adopted, starting with Science Citation Index. Since this covered only a small proportion of existing journals, it was accused of having a built-in bias; its criteria of "excellence" (based on citations) promoted English-language research at the expense of others, and also systematically excluded Third-World Science (6).

A problem of a different sort of iterated refinements in the measurement of quality, is that of incompatible indices. This also occurred in the science-policy field, in the definitions of an indicator for the quality of research institutes in West Germany (7,8). "Quality in Science" is no longer a debate among philosophers but a struggle for survival of the fittest when resources are scarce. In the German case, several indices, each quite plausible, of scientific quality (as, Ph.D's, staff acting as referees, job offers, and overseas visitors) were compiled; but the various rankings for quality among the institutions were all different! Any indicator would represent a policy choice external to the assessments made by the indices.

So far we have taken familiar examples to illustrate the complex way in which quality relates to indicators. Ecologists may derive some comfort and reassurance from this exposition; their problems of quantification are no worse in kind, although perhaps perceived as more severe in degree as well as very urgent. A discussion of quality leads necessarily to an analysis of the interpretation of "facts" and "values" in the definition of indicators. It might be argued that this gives ammunition to those who claim that it is all relative, a matter of rationalising commitments derived from power-politics and lifestyle (9). Our point, however, is that in the absence of such self-critical awareness, "statistics" are discredited (being popularly assigned to a class worse than "dammed lies"); and genuine debate, as well as informed policy-making, becomes very difficult to sustain.

An example of the effectiveness of constructive criticism in an ecological issue is the work of Michael Thompson on deforestation in the Himalayas. He first found that the only reliable quantitative fact about the situation is the extreme spread among the various official estimates of the indices used as crucial indicators of the problem. Thus, per-capita fuel-wood consumption varies by a factor of 67! However, this did not prove the unimportance of this indicator, only the misleading character of hyper-precise measures for the index. In fact, Thompson found universal consensus in the fact that the problem is indeed "serious"; and this non-quantitative (qualitative) scaling for the index is appropriate and, in policy terms, effective (10).

Should such non-quantitative indicators be thought "soft" or "non-scientific", we may recall all the non-quantitative indicators used in chemistry and the life sciences. Some appear quantitative, as the Apgar

Index, an indicator for a baby's condition at birth, where five attributes (colour, muscle-tone, response to stimulation, respiratory effort and heart-rate) are scored 0, 1 or 2 and added to provide an immediate assessment of the general health and viability of the baby (11). Purely qualitative indices are common in chemistry, as pH colour-scales. These may function merely as indices of the acidity of a solution, or (as when used in diabetics' urine) as an indicator of a potentially dangerous situation.

Finally, we come back to *quality of indicators,* in the sense of how well they perform their functions. Clearly, quality and uncertainty are in a sense opposed, but this is not a simple balancing. For uncertainty cannot be eliminated, and quality of indicators is gauged not so much by certainty of their forecasts as by justified confidence in their use. This is assured by a variety of means, including the ongoing processes of quality-control at all levels, down from definitions of indicators at the policy-level, through the technical work of construction of inidces, to the operational level of data-collection and anylysis. In these various ways, quality is assured and uncertainty is controlled. The quality of indicators is enhanced when the characteristic uncertainties can be managed and communicated as by NUSAP. Then policy-decisions utilizing such enriched indicators will be based in a more sophisticated analysis of options and the balances among them; in these, uncertainty, and even ignorance, are effectively brought into the equation.

References

(1) Funtowicz, S.O. and Ravetz, J.R., Policy-Related Research: a notational scheme for the expression of quantitative technical information, *Journal of the Operational Research Society,* 37, 1-5, (1986).
(2) Funtowicz, S.O., and Ravetz, J.R., Qualified Quantities - towards an arithmetic of real experience, In J. Forge (ed.), *Measurement, realism and objectivity,* 59-88, D. Reidel, (1987).
(3) Maddox, J., Half-truths make sense (almost), *Nature,* 326, 637, (1987).
(4) Mac Kay, A.L., *The Harvest of a Quiet Eye,* The Institute of Physics, Bristol and London, (1977).
(5) The index goes haywire, *The Guardian,* London (10/7/1985).
(6) Moravsik, M.J., Applied Scientometrics: An assessment methodology for developing countries, *Scientometrics,* 7, 165-176, (1985).
(7) Sietmann, R., School ranking inconclusive, *The Scientists,* (15/6/1987).
(8) Sietmann, R., West Germans debate research indicators, *The Scientist,* (29/6/1987).
(9) Douglas, M., and Wildavski, D., *Risk and Culture,* University of California, (1982).
(10) Thompson, M., Hartley, T., and Warburton, M., *Uncertainty on a Himalayan scale,* Sthnographica, London, (1986).
(11) Thomson, W.A.R., *Black's Medical Dictionary,* Adam and Charles Black, London, (1979).