Dealing with uncertainty in numbers

Jerry Ravetz [1]

[1]The Research Methods Consultancy Ltd. (RMC), London, United Kingdom

1. Why numerical uncertainty needs management

As of September 2001 the world suddenly became a very uncertain place. Armed with minimalist equipment and probably a minimal budget, a few dozen fanatics successfully waged war on the heartland of the United States of America. There was universal agreement that this was an event that defined the end of one epoch and the opening of a new one. What was last, particularly to Americans, was the sense of security and safety. Now there is a new sort of complicated war, against a new sort of covert enemy, in which uncertainty is at the core.

This new sensibility must affect all other areas of thought and practice. For the management of the environment, it will have a positive effect. The tendency to the complacency about the state of the environment, and the skepticism about warnings of environmental deterioration and harm, both widespread and officially sanctioned in America, will no longer feel plausible and right. Uncertainty will come to the focus of awareness; and its effective management in environmnental debates will be more important than ever.

In preparing for this new work, we must start with an appreciation that uncertainty is not well managed at present. In particular, number, the traditional language of science, is actually quite ill adapted for managing the sorts of uncertainty that we now confront. In this essay I will introduce a new system for the characterisation of uncertainty in quantitative information. Detailed discussions, and examples of fruitful application, are freely available elsewhere. Here I will content myself with showing that there is indeed a problem, that uncertainty exists and our existing conceptual tools are inadequate.

If not properly managed, uncertainty causes confusion in two ways. Frequently those who are promoting some development will cite numbers, frequently quite precise, which characterise both the benefits and the possible environmental costs on the basis of scientific studies. Those who have reservations about the project can voice their worries about ways that things can go wrong. But they will lose the rhetorical battle unless they can make a constructive criticism of the 'magic numbers' used to demonstrate safety, and can also express their concerns in a clear fashion. If their scientific case goes by default, they will then be handicapped in other aspects of the dispute as well.

It also happens sometimes that there is considerable uncertainty on all sides of a debate; this happens when global environmental issues are at stake. Then each side can 'play the numbers game', seizing on the uncertainties in each other's estimates. Without any agreed rules and discipline for managing the uncertainties, the game can go on without end. But since we have had no experiencing of evaluating scientific numbers, assessing their uncertainties and quality, we have great difficulty in bringing a discipline to those open-ended debates.

We now have a great and confusing contrast, between the precise scientific numbers we have always been shown, at school and by the experts, and the massive uncertainties that confront us when we try to engage in a real debate on environmental issues. Up to now, all our thinking about science has been based on the assumption that it produces hard, objective facts which are conveyed in precise numerical expressions. This is the lesson that students of science learn in their studies; and it is all the more persuasive because it is never mentioned. Students just learn that for every scientific problem there is a number, and just one number, that gives the answer. The idea that a scientific problem might have a whole range of numbers for the answer, or perhaps no numerical answer at all, just doesn't happen in science education. In this way we have actually been misled; in its emphasis on the great edifice of accomplished knowledge, our education has hampered our awareness of the important areas of ignorance that remain.

2. An example: hard decisions with soft numbers.

We are now coming to recognise that the science involved in environmental decision making is affected by deep uncertainties, for whch no answer may be forthcoming. For a current important example, there is the crucial parameter that comes into global climate change: the concentration of CO2 in the atmosphere. We have measures for it, direct and indirect; and we know that it has been rising steadily for some decades. We also have good reason to suspect that, acting as a 'greenhouse gas', CO2 will effectively trap more heat from the sun and produce a variety of changes in the processes that determine our climate. So we turn to the scientists for numerical answers to two big questions:

Is there a 'safe limit' of CO_2 concentration, below which there will not be serious effects of climate change?

How much do we need to reduce CO_2 emissions now in order to keep below that safe limit in the future?

There seems nothing wrong with such questions; those responsible for our health and safety calculate thousands of such 'safe limit' numbers every year. They are used to set standards and provide guidelines for the relevant inspectorates. But this 'safe limit' is different. It deals with a very complex system, the global climate. And it also deals with the future, representing conditions that are significantly different from those we can now measure and analyse. Computer models are used to simulate the interactions of the atmosphere with the biosphere and 'technosphere' of human activity. But the data is sparse and uneven in quality, the models are drastically simplified, calibration is complicated and indirect, and the extrapolations to the future decades ahead are quite insecure. Depending on the assumptions fed into the models, calculated consequences can vary between 'nothing serious' and 'disaster'. Some scenarios may be more likely than others; but none can be excluded.

In the absence of reliable information on the safe limit for CO_2, the argument over policy effectively becomes one of 'error-costs'. Those who worry about climate change concentrate on the worst case, and also the irreversible effects. The 'skeptics' concentrate on the costs to the economy of reductions in fossil fuel consumption, and stress the uncertainties in the CO_2 calculations. When policy recommendations are made, they tend to ignore the numbers and instead to convey their message by means of the cautions and caveats that qualify their advice.

After all the effort and resources that have gone into climate change science, mainly through the expensive computer simulations of the global climate systems, the numbers that they produce tend to be quietly pushed into the background when the real debate begins. But the numbers are not irrelevant. If their uncertainties and quality were properly understood, then they could be useful inputs. Suppose that it turns out that their uncertainty bounds are so very broad, and their quality in relation to the decisions is so low, that they are of little use as a basis for decisions. Well, then we learn something very important indeed: the depth of our ignorance. This could be one of the most important lessons that science can give us at the present time: to learn of our ignorance of what we are doing to the world around us. But so long as we believe that somewhere, somehow there is a magic number to be found which will define our problem and also its solution, we are fumbling in the dark. This is what we imbibe in traditional scientific education, and it is a lesson that needs to be unlearned without delay.

3. Uncertainty in numbers: when arithmetic crashes.

Of course, it is not so trivially simple. Scientists don't expect to dig out a single number like a nugget of fact from their experiments. When students do experiments in the lab they learn about the scattering of their data. If the student produces a report that has all the data points fitting exactly onto a trend line, the teacher will suspect that the data came from his calculator rather than from his equipment. But this sort of uncertainty is easily tamed. 'Error bars' can be calculated by simple arithmetical techniques; and then the final result can be expressed to the appropriate number of 'significant digits'. We call that sort of uncertainty 'technical', for it is managed by routine techniques. The success of those techniques in their proper realm gives the false impression, to the public and to many scientists alike, that all uncertainties can be tamed in the same way.

The public already has a critical awareness about one sort of numerical expressions. There is a well known saying about 'lies, damned lies and statistics'. Or as the Americans put it, figures can't lie, but liars can figure. The classic little handbook, *How to Lie with Statistics* (by Darrell Huff, Penguin Books, 1973+) gives a wealth of examples of how apparently objective statistics can be rigged. The techniques range over choice of sample, of question, and of representation, verbal and graphic. All this is now familiar fare. By contrast, we are still naive about the numbers themselves. Arithmetic is taught to young children, and so (we think) it must be elementary and simple. As a result we lack the realisation that there is a problem in the use of numbers; and if we don't know there is a problem we won't be thinking about a solution. Yet the problem is real, and has become acute in areas where there are great uncertainties, and where the quality of numerical data and expressions is up for question.

We might think of numbers and arithmetic as a very simple sort of computer. After all, computers are (in principle) just machines for doing lots of arithmetical operations very quickly. In their early days they were described as 'faster than thought' and 'unerring'. There was no theory of how computers could go wrong. But as they have come into universal used, we have learned that in spite of all their speed and great 'artificial intelligence', they can crash. How does a crash happen? It's when the instructions for the arithmetic can't be followed; instead of smoothly zipping along adding its 0's and 1's, the calculating 'brain' either goes into a closed loop or just jams. Could ordinary arithmetic crash too? We will see that it can, and does. Its rules can send misleading signals, and lead to confusion or error. When we see how even the simplest of arithmetical operations can 'crash' in its own way, we are ready to imagine the reasons why all numerical expressions, particularly in environmental issues, must be subjected to critical scrutiny.

To start with developing our critical awareness of numbers, in respect of their uncertainty and quality, there's no better way than a paradox. That shakes our complacency, and makes us think again about what we had always taken for granted. So we will tell a little joke, showing how our arithmetic can 'crash' when it is extended beyond calculation of sets of things, to the estimation of uncertain quantities.

The joke runs like this. A group of schoolchildren were visiting a Natural History museum, and an attendant was showing them a set of fossil bones. He told them that the fossils were 63,999,997 years old. Someone asked how he knew that, and he replied that when he started the job, he was told that the fossils were sixty-four million years old. But then he had heard that the age had been slightly overestimated; and so he had knocked off three years. To show his point, he did the sum on his clipboard:

64,000,000
-3
63,999,997.

This seemed bizarre. Where did all those 9's come from? Where there had previously been an easy string of zeros denoting the 'million', they had all suddenly flipped over and become quite precise 9's! Someone said, "That's ridiculous. Just a few years off sixty-four million doesn't make any difference; it's still sixty-four million". Annoyed, the attendant did another sum:

64,000,000
-3
64,000,000.

Then he showed his clipboard with the two sums to the class teacher. He challenged her to say which sum is correct. Is it the first one, that is arithmetically true but is practical nonsense; or is it the second one, realistic in practice but arithmetically wrong? If a pupil wrote down the second sum in answer to the question 'sixty-four million take away three', would the teacher give it a failing mark?

A great debate ensued. One student argued that when you say 'sixty-four million' you are counting in that sort of unit, of millions. Then smaller quantities are meaningless and the zeros for the smaller places are just there as a sort of filler. Another agreed, but asked how we know when a zero is a counter, or when it is a filler. Grabbing the clipboard, he wrote down a series of sums:

640
-3
637

6400
-3
6397

64,000
-3
63,097

640,000
-3
639,997

6,400,000
-3

Now, he asked, which of these is realistic? Which is describing a real counting, and which is more like the fossils joke? The first one might well be real, and the last two, with those strings of 9's in the answers, are obviously unreal; in the middle sums, the meaning of the zero depends on the context.

Some of the pupils became quite upset. Here was someone saying that 'zero' can mean two things, and that the meaning depends on the practical context. Why had noone had ever told them about this before! The teacher was also distressed. As far as she knew, arithmetic was a very simple subject, taught to the youngest children. Of course, there are odd things about 'zero', since you can't divide by it. But the idea that a digit, even zero, could be ambiguous in its meaning, was something new and unpleasant for her. It was as if her training in elementary mathematics had not prepared her, or her pupils, for the real world.

The fossils joke shows that numbers are used in a variety of ways, and that their meaning can depend on the context. When we are counting, the zero is really unambiguous. But when we are estimating uncertain quantities, it all depends. The six 'filler' zeros in the number for the age of the fossil are a way of saying that we are talking in millions, not in single years or in thousands of years. They are a means of expressing the degree of our uncertainty, or indeed our ignorance, about the precise quantity in question. But as we have seen, their meaning depends on conventions which are not merely unspoken, but which are generally unknown! In such conditions of uncertainty, the arithmetic we have learned in connection with simple counting and calculating is liable to crash.

While we are about it, we might ask about the 'sixty-four million' itself. The museum attendant interpreted a 'slight' difference in age as just three years; that would correspond to his scale of estimating time. But perhaps 'slight' meant 'a million', which is less than 2% of the sixty-four. And if the scientists are already calculating in estimates of millions of years, is there any certainty that 'sixty-four' is significantly different from 'sixty-three' or 'sixty-five'? Numbers have their own personalities: sixty-four is really telling us something about that last digit; if we said sixty-five, that might mean 'about halfway between sixty and seventy', about the right degree of precision. Perhaps the sixty-four really means 'nearly halfway between sixty and seventy'. And the apparently precise number is simply a code for a somewhat vague estimate.

The use of apparently hard numbers to express quite soft estimates happens all the time. A very familiar case is when precise numerical marks are given as quality assessments, which are usually based on subjective and contested judgments by the examiners. Now institutions are assigned places in 'league tables', with their associated rewards and punishments, which are defined by small differences between these precise numbers. We might think that the numbers that appear in scientific publications are very different from those produced by examiners, but that too depends on the context. In the old-fashioned physical sciences, they may well be more objective and 'hard'; but in the newer environmental sciences, where we are coping with uncertainty and ignorance, our numbers, reflecting our disagreements and judgments, are frequently quite soft. We need an arithmetic that reflects this new reality of policy and science, one that does not crash in the face of the new challenges and new sorts of decisions.

4. Describing uncertainty and quality in numbers.

'Quality' is one of the great growth industries of our time. Ever since the 1980's, when the Japanese suddenly swept the world market in automobiles by producing better cars at lower prices, we have been aware that 'quality' cannot be taken for granted. Consumers scrutinise products for information about their quality, and all sorts of institutions in the public and private sectors go through regular assessments of their quality. This effort has brought into being a vast array of organisations, public and private, which test for quality and instruct others on how to achieve it. Only one aspect of our lives is so far untouched by this general anxiety: numbers, or quantitative information in general. It is as if numbers are still accepted as nuggets of truth, simple and pure.

Perhaps this is one of the few surviving articles of faith in a generally post-modern age. We cannot predict how this particular lack of concern for quality would affect our society; it is not as if we eat numbers or use them to get from here to there. But since the quality of our decision-making, particularly on risks and the environment, is affected by our lack of awareness of quality in the scientific inputs, then we can be sure that sooner or later the harm will be real.

Changing this situation for the better will involve both a raising of awareness and also the provision of tools whereby the quality of scientific information can be assessed and then assured. Here will work on the latter task, hoping that it will also assist in the former. Our contribution is a notational scheme in which uncertainty and quality can be expressed in a concise and easily understood fashion. (For a full description, see *Uncertainty and Quality in Science for Policy*, by Silvio Funtowicz and myself; Kluwer, 1990). We call it 'NUSAP', the acronym being formed from the letters of the five categories of the scheme.

There is no problem in understanding the 'front end' of the notation, where the focus is more on uncertainty. The first three categories are Numeral, Unit and Spread. These correspond to the conventions in scientific writing. Thus if we have a measurement expressed as '30 km / hr ±.4', that is an instance of the first three categories in a common notation. For Numeral we have the ordinary numbers, for Unit we have kilometres per hour, and for Spread we have the 'error bar' ±.4. There is just one refinement that we should mention; the Unit category can (as in this case) be split in two, which we call the 'standard' and the 'multiplier'. This expresses the property of all measurements, of being referred back to a primary standard. In the case of length, it is a meter; and in the logical structure of the Système International, all other units of length are derived from it.

At this 'hard' end of the numerical expression, the principal use of the NUSAP scheme is to remind us of the flexibility we have in forming particular 'notations'. Thus the quantity in our example of Numeral might be expressed in many ways, including the 'scientific' notation 3E1. Or it might not be a number at all, but perhaps an ordinal (1st, 2nd,...). Awareness of the special character of Unit can help to clarify our thinking. The paradoxes of the 'fossils joke' can be partly resolved if we express the estimate of age as: 64 : Million-years . The multiplier 'Million' on the Standard 'year' makes it plain that we are estimating in millions of years, not in single years. Then we immediately see that the Spread announced by the museum attendant, giving 64 : Million-years : ±3 years is really quite meaningless. One might just as well have 64 : Million-years : ±3 milliseconds.

So far so good; now we consider the ways in which uncertainty merges into quality. We can start with the sorts of error that students encounter in lab experiments. We have already mentioned the 'scatter' which affects all data; this is conveyed by error-bars or our category of Spread. But there are other sources of error, most noticeably that of calibration. For a good example, we might imagine a sharpshooter aiming at a target. His shots might all cluster tightly around a particular spot, but if his gun sights are badly aligned that spot might be far from the bullseye! In this case, we can say that the sharpshooter achieved a really high precision, but unfortunately had a low accuracy. Experimental scientists have a convenient notation for expressing the accuracy, just adding another ± term. So we might see something like 30 km / hr ±.4 ±.6. This seems paradoxical, to have the inaccuracy greater than the imprecision! But if the scientists know from history that there is something still to be sorted out with their experiments, then that extra ± term is a prudent reminder of that deeper uncertainty.

Such a state of affairs is not uncommon; there is a classic case of a basic physical constant, the 'fine-structure constant', one of the 'dimensionless' numbers that define the structure of our physical universe.

[Figure 1 Here]

In the graph of the successive values of the quantity we see it going up and down between 1950 and 1968. We notice that the quantities are 'recommended values', arrived at by committes of experts, who would reconcile all the different values produced by different teams of experimenters. We see how the various values are expressed numerically with high precision, roughly one part in a million; but that the error bars extend over several parts per million at least. (The vertical scale on the left is used to indicate the magnitudes of the error bars). Worse, as we go from the earlier to the later recommended values, we see that more often than not the difference between successive values is greater than the Spread of any one of them. It is as if the scientists were constantly being surprised by the discovery of new sorts of error, which were producing inaccuracies that quite swamped the Spread of their previous estimates.

How would we use NUSAP to describe this basic physical constant, if we came at the end of that eighteen-year history of its bouncing through trial and error? Given that the deviation among the recommended values over that time was about twenty units (not counting the earliest value), then the last digit, representing a single part per million, is not meaningful. Our Numeral could then be 137.037 (to the nearest hundred-thousand), the Unit might be denoted ! to emphasise its dimensionless nature, and the Spread, composited from those on the graph, about 10 parts per million or one per hundred-thousand. And for the Assessment, it would only be prudent to say 1 at least, especially since the last value was less than the previous one by nearly 3 parts per hundred-thousand. In NUSAP, the estimated value for that whole historical development could then be: 137.037 : ! : ±1 : ±2. Even now the differences between estimates have not dwindled into insignificance; one Internet source quotes 137.03602855 ±1 (the reciprocal of .007297351 ±6), very close to the 1968 value but more precise by four (!) places; while another more cautiously gives 137.03597 ±2.

The sort of uncertainty conveyed by Assessment is not to be calculated routinely; the judgement depends on the experience of those with craftsmen's knowledge of their materials and tools. The category of Assessment need not be restricted to 'accuracy' as in the above example; it might be used to describe the reliability or 'strength' of scientific information in relation to its use in policy decisions. We might think of it as expressing a sort of 'safety factor', capable of giving a warning that too much of a load of argument is being placed on a single piece of evidence.

5. Numbers with Pedigree

Because the Assessment category is based on a judgment, there is no simple arithmetical routine that will produce a value for it. But it is possible to provide guidance for that judgment, and that is the function of the last category of the NUSAP scheme, Pedigree. With this category, we are definitely at the 'soft' end of the scheme, with a method for making informed judgments about quality. It might seem odd to imagine numbers having a pedigree, as if they were race horses or show dogs. But every number comes out of a process, partly technical with the materials and equipment, and partly social with the standards and conventions that guide the work. If we can characterise that process, then we are in a better position to define the limits of what it can accomplish. This can be of great assistance when we are trying to assess the strength of scientific information in relation to its place in a calculation or argument.

We might with that we had numbers that are precise to six digits and that have not changed in years of experiment or testing. But if we don't have them and can't have them in time for the decision, we must make do with what we have. There is no point in complaining that imprecise information on environmental parameters is 'inferior'; if it is the best we have, then it's the best, but we want to know how strong it is for our purposes. We might think of this judgment on the analogy of the classification of hotels. A five-star hotel will have more facilities than a two-star bed-and-breakfast; it is therefore in a higher class. But its quality may be affected by its context. If I am on a limited budget, its price may put it out of my reach, its luxuries are wasted on me, and I seek for the best that is available under the circumstances.

There is another dimension of quality, which it is important to realise cannot be expressed in the Pedigree method. Suppose that a particular luxury hotel is not all that it seems, and because of poor management its sanitary condition is below standard. Assessing this is a special task for the inspectors, who have a special competence; it will certainly not be revealed in the brochure! With the Pedigree method, we ascertain the 'best possible' quality within each class, in relation to the mode of production of the information and its intended uses. Digging deeper into a single item of infromation, amounts to a undertaking a refereeing operation, and that is not the business of this method.

For any class of problems and methods, there will be an appropriate form of pedigree. In each case, the pedigree is designed through a collaboration between experts in this sort of analysis and those who have a deep familiarity with the field. The task is to map the critical distinctions whereby judgments of quality are made. There is no 'correct' Pedigree; to make a useful Pedigree is a delicate design exercise. Returning to our first example, the modelling of global climate change, we show how the Pedigree can assist in the assessment of the scientific studies. The Pedigree method was applied, along with several other techniques, in a quality-assessment study of an important global climate model by a team led by Dr. J. van der Sluijs of the University of Utrecht, Netherlands. Full details of this research and on NUSAP itself can be found on the website http.www.nusap.net.

The usual form of the Pedigree is a matrix, or rectangular array. The columns represent the different aspects of the assessment. The cells in each column give the particular criteria, usually ranked in strength downwards from the top. Each row has a number, here running from 0 to 4, enabling a rough calculation to be made for the purposes of completing the Assessment.

Table 1 Pedigree matrix to assess parameter strength of climate models

Score	Proxy	Empirical	Theoretical understanding	Method	Validation
4	An exact measure of the desired quantity	Controlled experiments and large sample direct measurements	Well established theory	Best available practice in well established discipline	Compared with independent measurements of the same variable over long domain

3	Good fit or measure	Historical/field data uncontrolled experiments small sample direct measurements	Accepted theory with partial nature (in view of the phenomenon it describes)	Reliable method common within est. discipline Best available practice in immature discipline	Compared with independent measurements of closely related variable over shorter period
2	Well correlated but not measuring the same thing	Modelled/derived data Indirect measurements	Accepted theory with partial nature and limited consensus on reliability	Acceptable method but limited consensus on reliability	Measurements not independent proxy variable limited domain
1	Weak correlation but commonalities in measure	Educated guesses indirect approx. rule of thumb estimate	Preliminary theory	Preliminary methods unknown reliability	Weak and very indirect validation
0	Not correlated and not clearly related	Crude speculation	Crude speculation	No discernible rigour	No validation performed

An explanation of the elements is given in Appendix 1.

The Pedigree analysis has a number of applications. For potential users of the information, it provides a convenient account of the inherent strengths and weaknesses of the item of information. An evaluation that had previously been conducted among the experts and shared informally among them, inaccessible to outsiders, can now be public knowledge. This is an important contribution to the new policies of 'transparency and openness' which are to apply to decision-making on risks and the environment. Stakeholders from other areas of work will have a tool for analysis and constructive criticism; and this sort of empowerment will do a lot for making the new participatory processes effective and real.

But the Pedigree should not be seen only in terms of opening up secret knowledge for public scrutiny, a sort of bringing of a public gaze to the washing of scientists' dirty linen. Scientists themselves find it immensely beneficial. Quality control is not a regular, formalised activity among scientists in many fields. Aside from informal discussions among close colleagues, quality control is applied only prior to research, in peer review of applications, and after it, in refereeing of papers for journals. And in science unlike in most other areas, there is no structure of quality assurance, whereby the controllers are controlled. A notation like NUSAP, and the Pedigee analysis, will not by themselves accomplish the development of systems of regular quality control and quality assurance. But they provide the elements of the conceptual tools for the job. We can imagine that in every field there could be committees of scientists, assisted by NUSAP experts, designing and then applying the relevant Pedigrees. This has already started in the Netherlands, and can easily be extended elsewhere when the awareness and the will are there.

6. Conclusion.

This essay does not pretend to be an explanation of the NUSAP system; rather, it is an invitation to imagine the problem to which NUSAP offers a solution. Our whole scientific culture, over several centuries, has been built on the asssumption that science can provide assured knowledge, and that it is expressed in a precise numerical form. While that is still largely true on problems of science's own choosing, on problems of risks and the environment it can be a damaging illusion. I wanted to show that even on matters where scientific certainty seems unassailable, as in elementary arithmetic and fundamental physics, calculations can crash and constants can bounce. What then of the vast complex systems like the global environment?

On the great issues, like global warming, its effects and our responses, there will simply not be a scientific demonstration telling us what will happen and what we should do. Rather, there will necessarily be a dialogue among stakeholders and citizens. Science will provide essential inputs, but these will need to be assessed for their uncertainty and quality. NUSAP is a means of providing the tools whereby this can be done.

Appendix 1.

Proxy

Sometimes it is not possible to make a direct representation of the thing we are interested in by a numerical parameter, so some form of proxy measure is used. 'Proxy' refers to how good or close a measure of the quantity that we model is to the actual quantity we represent. Think of first order approximations, over-simplifications, idealisations, gaps in aggregation levels, differences in definitions, non representativeness, and incompleteness issues. If the parameter were an exact measure of the quantity, it would score four on proxy. If the parameter in the model is not clearly related to the phenomenon it represents, the score would be zero.

Empirical basis

The empirical basis typically refers to the degree to which the estimation of the parameter is based on direct observations, measurements and statistics. When the parameter is based upon good quality observational data, the pedigree score will be high. Sometimes directly observed data are not available and the parameter is estimated based on partial measurements or calculated from other quantities. Parameters determined by such indirect methods have a weaker empirical basis and will generally score lower than those based on direct observations.

Theoretical understanding

The parameter will have some basis in theoretical understanding of the phenomenon it represents. If our theoretical understanding of some mechanism is very high, we may well be able to make reliable estimates for the parameters that represent that mechanism, even if the empirical basis is weak. On the other hand a strong empirical basis may not be sufficient to estimate future values of parameters if our theoretical understanding of the mechanisms involved is absent. In that case, extrapolation from past data is not warranted. This criterion aims to measure the depth of the theoretical understanding that was used to generate the numeral of that parameter. Parameters based on well-established theory will score high on this metric, while parameters whose theoretical basis has the status of speculation will score low.

Methodological rigour

Some methods will be used to collect, check, and revise the data used for making parameter estimates. Methodological quality refers to the norms for methodological rigour in this process that are applied by peers in the relevant disciplines. Well-established and respected methods for measuring and processing the data would score high on this metric, while untested or unreliable methods would tend to score lower.

Validation

This metric refers to the degree to which one has been able to cross-check the data and assumptions used to produce the numeral of the parameter against independent sources. Since climate models cannot be tested in the ordinary way, their 'validation' must be indirect. We can assess the degree to which the assumptions used to produce the numerical parameter have been cross-checked against independent sources. In many cases, independent data for the same parameter over the same time period are not available and other data sets must be used for validation. This may require a compromise in the length or overlap of the data sets, or may require use of a related, but different, proxy variable for indirect validation, or perhaps use of data that has been aggregated on different scales. The more indirect or incomplete the validation, the lower it will score on this metric.

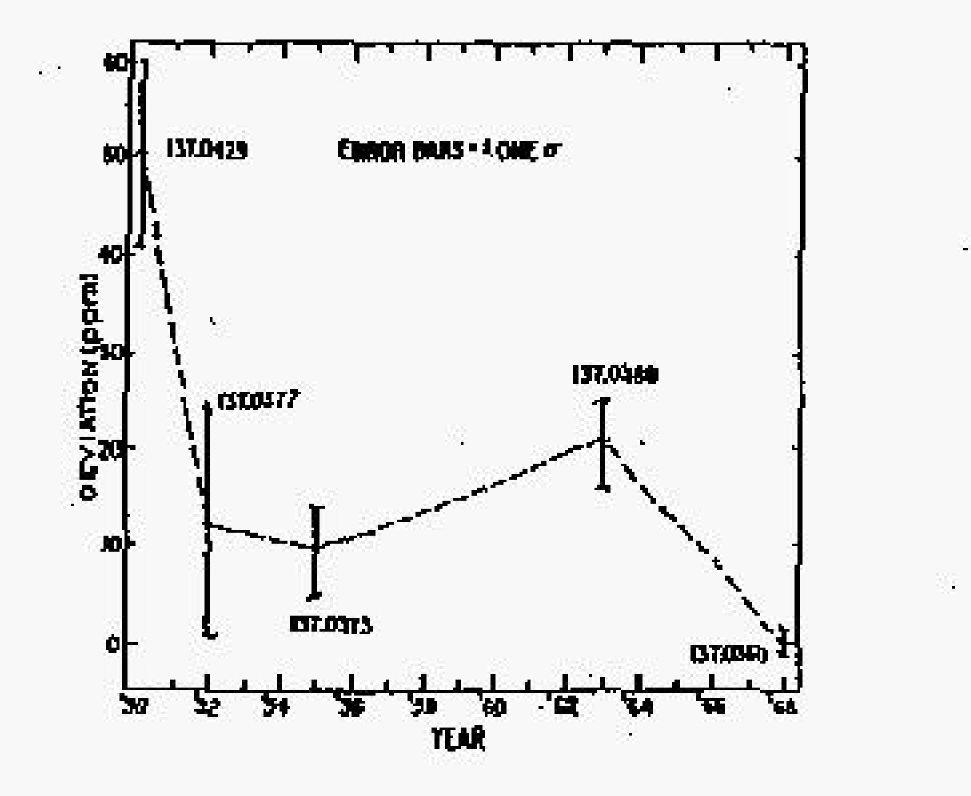

Figure 1. Successive 'recommended values' of the fine-structure constant