Calls for the closer integration of science in political decision-making have been commonplace for decades. However, there are serious problems in the application of science to policy — from energy to health and environment to education.

One suggestion to improve matters is to encourage more scientists to get involved in politics. Although laudable, it is unrealistic to expect substantially increased political involvement from scientists. Another proposal is to expand the role of chief scientific advisers, increasing their number, availability and participation in political processes. Neither approach deals with the core problem of scientific ignorance among many who vote in parliaments.

Perhaps we could teach science to politicians? It is an attractive idea, but which busy politician has sufficient time? In practice, policy-makers almost never read scientific papers or books. The research relevant to the topic of the day — for example, mitochondrial replacement, bovine tuberculosis or nuclear-waste disposal — is interpreted for them by advisers or external advocates. And there is rarely, if ever, a beautifully designed double-blind, randomized, replicated, controlled experiment with a large sample size and unambiguous conclusion that tackles the exact policy issue.

In this context, we suggest that the immediate priority is to improve policy-makers’ understanding of the imperfect nature of science. The essential skills are to be able to intelligently interrogate experts and advisers, and to understand the quality, limitations and biases of evidence. We term these interpretive scientific skills. These skills are more accessible than those required to understand the fundamental science itself, and can form part of the broad skill set of most politicians.

To this end, we suggest 20 concepts that should be part of the education of civil servants, politicians, policy advisers and journalists — and anyone else who may have to interact with science or scientists. Politicians with a healthy scepticism of scientific advocates might simply prefer to arm themselves with this critical set of knowledge.

We are not so naive as to believe that improved policy decisions will automatically follow. We are fully aware that scientific judgement itself is value-laden, and that bias and context are integral to how data are collected and interpreted. What we offer is a simple list of ideas that could help decision-makers to parse how evidence can contribute to a decision, and potentially to avoid undue influence by those with vested interests. The harder part — the social acceptability of different policies — remains in the hands of politicians and the broader political process.

Of course, others will have slightly different lists. Our point is that a wider

**Twenty tips for interpreting scientific claims**

This list will help non-scientists to interrogate advisers and to grasp the limitations of evidence, say William J. Sutherland, David Spiegelhalter and Mark A. Burgman.
Practically all tends to give an exaggerated picture of the significant results are more likely to be written up and published, the scientific literature will sample different proportions of the population, and all may give different results.

Bias is rife. Experimental design or measurement processes may produce atypical results in a given direction. For example, determining voting behaviour by asking people on the street, at home or through the Internet will sample different proportions of the population, and all may give different results.

Bigger is usually better for sample size. The average taken from a large number of observations will usually be more informative than the average taken from a smaller number of observations. That is, as we accumulate evidence, our knowledge improves. This is especially important when studies are clouded by substantial amounts of natural variation and measurement error. Thus, the effectiveness of a drug treatment will vary naturally between subjects. Its average efficacy can be more reliably and accurately estimated from a trial with tens of thousands of participants than from one with hundreds.

Correlation does not imply causation. It is tempting to assume that one pattern causes another. However, the correlation might be coincidental, or it might be a result of both patterns being caused by a third factor — a ‘confounding’ or ‘lurking’ variable. For example, ecologists at one time believed that poisonous algae were killing fish in estuaries; it turned out that the algae grew where fish died. The algae did not cause the deaths.

Regression to the mean can mislead. Extreme patterns in data are likely to be, at least in part, anomalies attributable to chance or error. The next count is likely to be less extreme. For example, if speed cameras are placed where there has been a spate of accidents, any reduction in the accident rate cannot be attributed to the camera; a reduction would probably have happened anyway.

Extrapolating beyond the data is risky. Patterns found within a given range do not necessarily apply outside that range. Thus, it is very difficult to predict the response of ecological systems to climate change, when the rate of change is faster than has been experienced in the evolutionary history of existing species, and when the weather extremes may be entirely new.

Beware the base-rate fallacy. The ability of an imperfect test to identify a condition depends upon the likelihood of that condition occurring (the base rate). For example, a person might have a blood test that is ‘99% accurate’ for a rare disease and test positive, yet they might be unlikely to have the disease. If 10,001 people have the test, of whom just one has the disease, that person will almost certainly have a positive test, but so too will a further 100 people (1%) even though they do not have the disease. This type of calculation is valuable when considering any screening procedure, say for terrorists at airports.

Controls are important. A control group is dealt with in exactly the same way as the experimental group, except that the treatment is not applied. Without a control, it is difficult to determine whether a given treatment really had an effect. The control helps researchers to be reasonably sure that there
Separate no effect from non-significance. The lack of a statistically significant result (say a P-value > 0.05) does not mean that there was no underlying effect; it means that no effect was detected. A small study may not have the power to detect a real difference. For example, tests of cotton and potato crops that were genetically modified to produce a toxin to protect them from damaging insects suggested that there were no adverse effects on beneficial insects such as pollinators. Yet none of the experiments had large enough sample sizes to detect impacts on beneficial species had there been any.

Effect size matters. Small responses are less likely to be detected. A study with many replications might result in a statistically significant result but have a small effect size (and so, perhaps, be unimportant). The importance of an effect size is a biological, physical or social question, and not a statistical one. In the 1990s, the editor of the US journal Epidemiology asked authors to stop using statistical significance in submitted manuscripts because authors were routinely misinterpreting the meaning of significance tests, resulting in ineffective or misguided recommendations for public-health policy.

“The question to ask is: ‘What am I not being told?’”

Study relevance limits generalizations. The relevance of a study depends on how much the conditions under which it is done resemble the conditions of the issue under consideration. For example, there are limits to the generalizations that one can make from animal or laboratory experiments to humans.

Feelings influence risk perception. Broadly, risk can be thought of as the likelihood of an event occurring in some time frame, multiplied by the consequences should the event occur. People's risk perception is influenced disproportionately by many things, including the rarity of the event, how much control they believe they have, the adverseness of the outcomes, and whether the risk is voluntarily or not. For example, people in the United States underestimate the risks associated with having a handgun at home by 100-fold, and overestimate the risks of living close to a nuclear reactor by 10-fold.

Dependencies change the risks. It is possible to calculate the consequences of individual events, such as an extreme tide, heavy rainfall and key workers being absent. However, if the events are interrelated, (for example a storm causes a high tide, or heavy rain prevents workers from accessing the site) then the probability of their co-occurrence is much higher than might be expected. The assurance by credit-rating agencies that groups of subprime mortgages had an exceedingly low risk of defaulting together was a major element in the 2008 collapse of the credit markets.

Data can be dredged or cherry picked. Evidence can be arranged to support one point of view. To interpret an apparent association between consumption of yoghurt during pregnancy and subsequent asthma in offspring, one would need to know whether the authors set out to test this sole hypothesis, or happened across this finding in a huge data set. By contrast, the evidence for the Higgs boson specifically accounted for how hard researchers had to look for it — the ‘look-elsewhere effect’. The question to ask is: ‘What am I not being told?’

Extreme measurements may mislead. Any collation of measures (the effectiveness of a given school, say) will show variability owing to differences in innate ability (teacher competence), plus sampling (children might by chance be an atypical sample with complications), plus bias (the school might be in an area where people are unusually unhealthy), plus measurement error (outcomes might be measured in different ways for different schools). However, the resulting variation is typically interpreted only as differences in innate ability, ignoring the other sources. This becomes problematic with statements describing an extreme outcome (‘the pass rate doubled’) or comparing the magnitude of the extreme with the mean (‘the pass rate in school x is three times the national average’) or the range (‘there is an x-fold difference between the highest- and lowest-performing schools’). League tables, in particular, are rarely reliable summaries of performance.

William J. Sutherland is professor of conservation biology in the Department of Zoology, University of Cambridge, UK.

David Spiegelhalter is at the Centre for Mathematical Sciences, University of Cambridge. Mark Burgman is at the Centre for Excellence in Biosecurity Risk Analysis, School of Botany, University of Melbourne, Parkville, Australia.

e-mail: wjs32@cam.ac.uk