Interpreting science – implications for public understanding, advocacy and policy formation

A discussion paper

Sir Peter Gluckman KNZM FMedSci FRSNZ FRS
Chief Science Advisor to the Prime Minister

April 2013

Office of the Prime Minister's Science Advisory Committee
PO Box 108-117, Symonds Street, Auckland 1150, New Zealand
Telephone: +64 9 923 1788
Website: www.pmcsa.org.nz
Email: csa@pmcsa.org.nz

Preface

Society uses science in many ways for its benefit – and sometimes individuals and groups within society can misuse science either accidentally or intentionally. However, the proper use of science and technology is essential to our economic, social, and environmental health. As science and technology are being used to address increasingly complex issues and policy makers face the difficult choices on how to reach trade-offs between contesting views and inputs, science has become more important in providing a relatively value-free knowledge base on which the public and policy maker can reflect, integrate their own values and priorities, and make decisions to use or limit technologies or to introduce new policies and programmes or to change current ones.

Thus how society obtains and understands scientific and technical knowledge is critical to a well performing participatory democracy. But because science now deals with very complex matters, many of which have high values content (for example environmental issues, the development of behaviour, the use of genetic modification), how that science is presented and used can have major impacts on societal decision-making and progress.

Clearly those who are active in science communication have major roles to play in allowing the public and the policy maker to understand what we know, what we do not know, and the nature of the inferential gap between what we know and what might be concluded.

Too often a piece of science is misunderstood, misused or overstated – sometimes something is presented as established science when it is not, other times it does not suit advocates to accept the science as established when it is. This paper will give examples of each of these and highlight the questions that should be asked when interpreting evidence. It also explains how scientific conclusions can be established even when all the details may never be resolved or there is still debate over some specifics; classic but diverse examples of this are our understandings of evolution and of earthquakes.

Two matters have given me particular concern. The first has been the increasing trend for the complex nature of science to be ignored or misunderstood. It leads to the – sometimes rhetorically convenient – argument that you can find a scientist to support any given position. This totally misinterprets the scientific endeavour and does grave mis-service to the public interest. Scientific consensus is unlike social consensus – it is not a matter of the loudest voice or compromise; it is a more consultative process by which the expert community examines the currently available evidence and incorporates it into an understanding that integrates what we know and acknowledges what we do not know. The very nature of the natural world and scientific observation means that variable results can often be expected and the process of scientific consensus addresses this problem.

The second challenge is that of science being wrongly used as a proxy for a debate over values. This may occur consciously or unconsciously. It is obviously psychologically easier in some situations to say the science is not settled, or not settled enough, than to enter more complex discussions that have strong values components. Much of the climate change debate has used science as a proxy when the real debate, which is valid, is over responsibility between nations and between generations.

The miscommunication and misuse of science in the public domain can engender serious mistrust in the scientific enterprise. The scientific community has to do much to improve its behaviour, but equally society will better served if science is not miscommunicated and not misused in advocacy or policy formation. We do not live in a technocratic society, we live in a democracy and values will always and should be the final arbiter of decisions that are made. Values are formed and moulded by what we know and what we think. We often interpret what we hear or learn in the light of pre-existing biases. Science is the only process we have to gather reliable knowledge about our world, our society and our environment. It should therefore be seen as an essential input into the formation of opinion and values, but it can only do that if its is honestly represented and honestly used, and if society is provided with adequate opportunity to understand the scientific process.

The challenges of the twenty-first century are many – whether we look at the obvious issues of climate change, sustainable economic growth and resource security or the more subtle issues arising from greater urbanisation, changed ways of communicating and the changed nature of our society, science and technology will be essential to navigating a productive and safe path to the next generations. If science is not used and communicated in a way that is appropriate and with high integrity and fidelity we risk sailing into dangerous waters. This paper is intended to assist in better navigation by highlighting how to interpret complex science.

Sir Peter Gluckman

April 2013

1. Introduction

In 2011, I released a discussion paper entitled *Towards better use of evidence in policy formation* which discussed the interaction between knowledge and policy making. It points out that the process of policy formation is improved if evidence is first incorporated in a value-free manner and only then should the various values-laden domains such as public opinion, fiscal priorities, diplomatic concerns and electoral considerations be overlaid upon knowledge. When the science itself is presented in a values-laden way it is compromised and loses its privileged place in policy formation. Conversely, the failure to use evidence properly can lead to decision making which is less likely to produce effective and efficient outcomes.

Beyond the obvious domain of ethics, science is never absolutely value-free. The key values domains to consider are, first, expert judgements over the quality and sufficiency of data and, second, the limits of the data available. There are nearly always inferential gaps between what is known and the decisions that are implied by the knowledge. The gaps and uncertainties must be acknowledged. If that is done with integrity then science advice can be delivered in an effectively value-free way. In *Towards better use of evidence in policy formation* I argued that science should be presented in a manner that is not based on advocacy but is delivered by 'honest brokerage' to the policy maker. It is for the policy maker to overlay the other critical domains of policy formation.

We live in a participatory democracy and ultimately it is for the public and the policy maker to use scientific knowledge – the challenge is how to assist them to use it properly. It is therefore important that citizens and policy makers become aware of the uses and misuses of scientific data. Often it is easy to find an apparent scientific claim or data to support any particular intervention or action, or argue for or against a particular policy change – but is that data sufficiently sound or reliable to act as a basis for decision making? There are many traps in extrapolating from a single study, and this is an important reason why expert but impartial advice is needed in bridging science and policy.

A further issue emerges because the media often inappropriately interprets or profiles a particular single report or claim because it will attract attention and have impact, and it is this over-emphasised and sometimes contrary claim that influences public opinion, even in the face of considerable opposing evidence. A more engaged and higher quality of communication between scientists and society is essential if society is to make better use of scientific knowledge.

This discussion paper is intended to highlight some of the matters that need to be considered when interpreting science.

2. The nature of the scientific process

The science section of the *New Zealand Curriculum* rightly focuses on teaching the 'Nature of Science', emphasising to students that science is an organised process for obtaining new knowledge – not simply a collection of facts. There are many approaches to science but in general they involve making observations, using inference to generate a hypothesis that explains the observations, gathering evidence to support or refute the hypothesis and modifying it as needed to accommodate the new evidence, and testing the hypothesis by seeing whether it predicts new facts.

Critically, there must be a high level of integrity in analysing and interpreting the data. This is protected to a considerable extent by two other important foundations – replication (where repeating a study gives confidence as to its reliability, so scientific papers usually provide careful description of the methodology so that other qualified researchers can repeat the work) and expert peer review (evaluation of studies by other qualified individuals so as to maintain quality and consistency). Here scientific journals have a central role in providing a mechanism for peer review and a vehicle for registering and disseminating research findings. The scientific community generally does not accept claims made in the absence of expert peer review.

3. Does a study show what it claims?

Publication of a peer-reviewed paper in a reputable journal is considered a generally reliable – but not infallible – mark of scientific quality. So perhaps the first question for policy makers when interpreting a scientific claim should be – has it been published? Nevertheless, science is a human activity and scientists are subject to the same career pressures as in any other profession. Just because a paper is published does not mean that it is not flawed. Some of the causes of such flaws are discussed below. Equally, scientists have their own personal values that may affect their interpretation of their work and the advice they give. Sometimes this can lead to tragic consequences (see Box 1). While these are exceptions, they highlight the danger of reliance on single or

Box 1: Vaccine safety falsely questioned

In 1998, Andrew Wakefield and his collaborators published in *The Lancet* a paper claiming to show a link between administration of the measles, mumps and rubella (MMR) childhood vaccine and the subsequent development of autism and bowel disease. The study was widely publicised, and the resulting decrease in vaccination rates led to outbreaks of measles, resulting in deaths and permanent disability. But other researchers failed to confirm the study's findings and it received widespread scientific criticism because of the small number of children studied, the selection bias in recruiting them, and the methods used. The research was later shown to be fraudulent and the paper was retracted.

Although this unfortunate episode represents a failure of peer review and editorial judgement, in the end the self-correcting nature of the scientific process ensured that the false claims did not persist in the scientific community.

extreme studies. In the case of the MMR vaccine, echoes still persist and are seized on by anti-vaccine advocates.

Given the competition for funding and the natural human desire for recognition, it is perhaps understandable that too many scientists, aided and abetted by the communications departments of their institutions, are tempted to overstate the implications of their work – cures for cancer or pronouncements on the human condition appear regularly in stories in the media. Similarly, values-driven pressure groups will 'cherry-pick' studies that they can present as credible and convincing to support their particular advocacy agenda, however weak the findings might appear to the informed expert, .

The self-correcting nature of the scientific process – hypothesis modification, replication and peer review – will in time ensure the probable correctness of the scientific record – but in the short term, how should policy makers and concerned citizens assess the validity of, and the claims made for, any single study? This requires greater engagement between scientists and society through which the implications of the knowledge can be discussed, interpretation of evidence clarified, and applicability to a particular context explored.

First it must be acknowledged that technical problems around study design or implementation can cause problems that only expert commentary or investigation can detect. We saw an example in the announcement in 2011 of neutrinos that appeared to travel faster than light – apparently breaking a fundamental law of physics. The well-publicised paper caused a storm of criticism and commentary, and in trying to replicate their results the researchers found a loose wire in their apparatus that affected their timing measurements. As the research director of the European Organization for Nuclear Research, one of the institutions involved, put it:

> *"The story captured the public imagination, and has given people the opportunity to see the scientific method in action – an unexpected result was put up for scrutiny, thoroughly investigated and resolved in part thanks to collaboration between normally competing experiments. That's how science moves forward."*

This episode nicely illustrates several points mentioned above – the increasing pressure to publish exciting results, the importance of replication, and the self-correcting nature of science. Critiquing the validity of that paper clearly needed technical expertise, but below we discuss some other ways in which non-technical readers, both public and policy makers, can begin to judge scientific claims.

4. Variability

One of the biggest issues is that of variability in reported results and the conclusions reached about the same problem. There are many examples where there is confusion – does drug education in high schools prevent drug use or not? Exact replication is nearly always difficult in policy-relevant social studies and indeed in biological and medical studies. Even assuming the same general approach was used there can be an enormous variation in the claimed results because of differences in study populations and the details of the intervention.

Smaller studies are particularly vulnerable in this way because of the inherent problems in making conclusions from minimal data that do not represent a valid sample of the population. It is no different to assuming that the average height of New Zealanders is 185 cm just because the first ten people you might have seen when you arrive in New Zealand happen to be members of a basketball team. The smaller a sample or the smaller the experiment, the less representative it is. That is why smaller sized opinion polls have greater margins of error. Increasing the sample size ought to increase the representativeness of the sample, but unless the randomness of the sample selected has itself been assured by some statistical process, then the reliability of inferences made about the whole population will be lost. An example might be opinion polls carried out by random calling of fixed domestic telephones: increasing the number of calls will increase the reliability of the conclusions reached from that sample, but does not overcome the problem that people answering domestic telephones might not be representative of the whole population, for instance because younger people tend to have mobile phones rather than land lines.

The implications can be enormous. One study might find that an intervention works and another study does not (for example, in evaluating whether drug education in high school works). This variation leads to a real danger of 'cherry picking', where an advocate will emphasize one particular study and its results because it confirms a pre-existing bias. These issues are arising more frequently because the nature of questions science now engages with are more complex and often relate to social, environmental and human matters. Thus confusion can be accidentally or wilfully created, leading to widespread confusion. When confusion is widespread, then belief and dogma become the sole basis of decision-making with the risks that that must entail.

Science has developed ways to address these issues of variability. The key issue is deciding whether the number of subjects or objects being studied is sufficient. This can be done by calculating the 'power' of a study – that is the number of subjects necessary to detect an effect of a pre-specified size – before the study is actually started. The nature of statistics is such, that if sufficient comparisons are made some associations will be found through chance alone and in reality these are not true effects. It is therefore desirable that scientists pre-specify the hypothesis being tested. This means first declaring what is the question is that needs to be asked and then set up the study appropriately to test that question. If it is not done in that order, gross errors can be made.

There are statistical approaches to combine independent studies addressing the same question to overcome the variability that individual studies may have, depending on the domain being studied. This is known as 'meta-analysis' (see Box 2). It too has its limitations in that its value depends on how many studies have been included or excluded on the basis of quality and methodological variability. The choice of these various approaches is in itself a matter of particular expertise.

5. Association, causation and confounding

There is another fundamental problem that frequently is misused by advocates and indeed misunderstood by many – the meaning of an association or correlation. Just because one thing is statistically associated with another does not imply that the first factor causes or influences the second factor – they may in fact have no relationship. It may look like a plausible relationship, but such correlations do not mean that causal relationships are indeed present. The second factor may actually cause the first factor, or it could be that a third factor (a confounder) influences both the first and second factors independently and the association is entirely irrelevant. In many social and medical questions, socioeconomic factors are an important confounder and unless that is properly controlled for, studies with populations of different socioeconomic status may lead to different effect sizes or conclusions.

An over-simplified example may help to understand confounding: over the course of a year we would probably see a temporal association between ice-cream consumption (first factor) and cases of sunburn (second factor). So should we deduce that eating ice-cream causes sunburn? Obviously not – eating ice cream does not cause sunburn, but rather summer (the confounder) is independently associated with more cases of sunburn and more ice cream eating.

Box 2: Meta-analysis

Meta-analysis is a formal methodology that can integrate different studies of various quality, size and conclusion and produce an estimate of the most likely overall effect and its level of certainty. It has most often been used when determining the effects of medical or public health interventions, but can be applied to other areas. Meta-analysis combines the results of studies of the same research question using carefully designed statistical procedures that ensure that any single study does not unduly bias the combined outcome of the analysis.

The best established use of meta-analysis is the work of the *Cochrane Collaboration*, which is a network of 28,000 researchers from more than 100 countries who use meta-analysis and related techniques to promote evidence-based health care.

Box 3: Smoking and cancer

British epidemiologist Sir Richard Doll is best remembered for demonstrating the link between smoking and lung cancer, and his work provides an excellent example of establishing a causal relationship between an environmental factor and a disease. His first study published in 1950 involved analysis of the smoking habits of individuals with and without lung cancer. Such **case-control studies**, where the characteristics of people who have a disease are compared with those of similar people who do not, so as to identify contributing factors, are powerful research tools in medicine – but resolving causation requires more. Doll noted the correlation between smoking history and development of the disease but, correctly, restricted itself to a claim that "smoking is an important factor" (although Doll himself was sufficiently convinced that he gave up smoking). A later paper published in 1954 reporting the smoking habits of individual British doctors and their subsequent mortality from lung cancer (a **cohort study**) strengthened the correlation but was still not enough to prove causation. A later **intervention study**, in which one group in the community (doctors) was persuaded to stop or reduce smoking to a greater extent than the population as a whole, with a resulting decline in their mortality from lung cancer relative to the population, confirmed the role of smoking in the causation of lung cancer.

A more serious example arises from Wakefield's work (see Box 1), where it was claimed that vaccination caused autism because there was some correlation between introduction of new forms of vaccination and the increasing number of cases of autism diagnosed. The connection appeared superficially plausible and was believed by many, but careful research showed this was not a real association at all. Simply, the two events were concurrent with time – that is, at roughly the time the new vaccinations were introduced, the diagnostic criteria for autism were widened and the methods for the diagnosis of autism were improved. Time was effectively the third factor and the two primary factors, autism and vaccination, were not related.

A further example: the apparent incidence of child abuse is increasing. Is this real or are we just getting better at detecting and reporting child abuse? If it is real, can we relate it to changing social circumstances (say, more single parent families) simply because over the same timeframe family structures have changed? To be extreme with an absurd hypothetical, it could be caused by the greater use of mobile phones that has occurred in the same time frame. The former may be more plausible, but the association does not prove causation. More complex analyses are needed.

In the end it can only be experiments or interventions that can prove associations to be causations. To go back to our first example, we could give people ice-cream and see whether they get more sunburn – only then could we resolve whether this is a meaningless association or a causative relationship (see also Box 3). In many cases, however, such studies are impractical or unrealistic and indirect means are needed to decide whether a statistical association implies a causal relationship or not.

There is another trap in looking at what happens over time and whether an intervention has been effective or not. This is technically known as the 'regression to the mean' – it refers to the situation where if something measured is much worse (or much better) than the average value on the first observation, it is likely to be closer to the average, and therefore apparently improved (or deteriorated), on subsequent observation. For example, a claim that a particular road safety initiative introduced after a spate of accidents has reduced the number of fatalities may be an instance of regression to the mean rather than a true causal relationship.

Sometimes however the choice to intervene must be based on only a plausible association and not on definitive interventional experiments because there is no other way to progress. In those cases, evaluation is critical – otherwise, investment may continue in a meaningless, even if apparently well justified, intervention.

It is unfortunate that much public expenditure is on programmes that may or may not be effective, because evaluation has been lacking. This lack of evaluative discipline particularly in relationship to social interventions means that much in social and public policy becomes entirely based on opinion and anecdote rather than on evidence – and irrespective of the desirability of particular outcomes, they are less likely to be assured. While pilot studies and/or controlled trials of a particular intervention (for example altered class size) may be seen as slowing policy development or in some cases create controversy (because some individuals are seen as missing out or being used as guinea pigs), the greater use of such scientific approaches to policy development must lead to more consistent and valued outcomes.

6. Statistics and measurement

Statistical tests are the tools used to deal with the problems of inherent variation and chance. Many things can change by chance, and statistical tests are used to see if an association is likely to be real rather than coincidental. The choice of appropriate statistics is complex and highly technical and bad statistical analysis can be quite misleading – again this is a matter for unbiased expertise.

Generally scientists address these problems firstly by good experimental design, secondly by using larger and properly randomised samples (making the risk of a statistically erroneous solution less likely) and thirdly and most importantly by replication of a study. Repeatability is the major tool to seek validity of a conclusion.

Underlying all of this is the need to measure something that is meaningful. A failure to have a meaningful and accurate measure can mean all sorts of errors can arise. Thus in outcomes based research it is very important to be sure what is the outcome being measured, that it really represents the goal we want to achieve, and that it can be measured reliably. For example, enthusiasts for a particular preschool initiative might pronounce it a success on the basis of parent satisfaction (which might simply reflect the child-care element of the programme) rather than measuring the outcomes of real value to the child and interest to society, such as the child's emotional or cognitive development.

Sometimes, though, for practical reasons we have to measure indirect markers of an effect (surrogate measures) rather than important real-world outcomes. For example, if it is known from earlier studies that people with lower blood cholesterol levels have fewer heart attacks and so live longer, then a dietary intervention might be approved on the basis of its ability to lower blood cholesterol (the surrogate measure) rather than its ability to reduce heart attacks and increase lifespan (which are the primary concerns), and which would take a very long period of observation to ascertain.

7. Absolute and relative effects

Even if one factor does cause a consequential change in a second factor, there is a much bigger issue that is frequently forgotten by advocates and yet is key to the decision-making processes of the policy maker. Scientific assessment of a result needs to consider not only if the first factor causes change in the second factor but *how much* of a change it causes. So when looking at risk or benefit, we must think about the absolute effect rather than simply proportionate or relative effects – this is what is referred to as effect size.

Imagine that exposure to a particular chemical increases the risk of a very rare disease, say from 1 in 1,000,000 people to 1 in 500,000 people after 10 years exposure. In proportionate terms this is a doubling of, or 100% increase in, the risk and one can imagine the dramatic headlines that might emerge or the hyperbolic claims an advocate for banning the chemical might make. But considering the same data in terms of absolute risk shows the effect to be trivial from a population perspective. Even if every New Zealander was exposed for a decade, the number of additional cases caused by the presence of the chemical would be only four or five. Now consider at this same data from the public policy perspective. Not only does the absolute risk matter but also does the nature of the risk, which might range from a mild rash easily treated to a serious cancer or birth defect. The policy response might thus depend not only on the change in disease risk but severity of the outcome under consideration. The policy maker also has to consider the costs of removing the chemical from our ecosystem (it might be something of no fundamental importance which can be easily removed or it might be essential to a major industry). The balance of public policy may well decide that there is no justification for action to remove the chemical. Yet an advocate for its removal may claim that the chemical has 'doubled the risk' of disease – scientifically true but inherently misleading. The absolute risk is of rather less consequence for citizens and policy makers. The policy framework would however be very different if the baseline disease incidence was 1 in 1000 people and the absolute risk then doubled to 1 in 500 – the number of New Zealanders affected would be in the thousands and even if the chemical was core to a major industry and even if the outcome was only one of a skin rash, action would really matter.

This issue of absolute versus relative effect is critical to the policy formulation process. Not infrequently we see over-reaction and over-claims made by advocates in relationship to very small absolute effects. Given that across all policy making, different options for the application of the limited funds available to society may exist, this dimension is a key consideration.

8. Context

With any intervention or action we need to understand its context, which in turn impacts on the effect size. Consider breast-feeding. While there are many well-proven advantages to breast-feeding, its most important benefit is that it provides significant immunity from infection before babies fully develop their own immune systems. This led to the critical drive to promote breast-feeding in the developing world, where failure to breast-feed may well expose the infant to devastating and even fatal infection. Yet in the modern hygienic world of developed countries, the effect of breast-feeding on resistance to infection is of minor importance compared to the effects of good hygiene, immunization and adequate nutrition. In a developed country setting, breast feeding perhaps reduces the frequency of colds by one episode in the first year of life. The effect sizes and arguments for breast-feeding in the least developed world and the developed world are thus very different. There are of course other benefits of breast-feeding: these include improved emotional attachment and possibly better brain function and less obesity for the child. Again, however, there are other considerations – if the primary argument for promoting breast-feeding is about attachment and if the pressures to breast-feed are too zealous and affect the mother's emotions or social capacities, the effort might be counterproductive.

In this example, as in many others (see also Box 4), there may be evidence for many different effects, but the key scientific questions which advocates can lose sight of are – what is the effect size and in what context? Policy makers have to make trade-offs between many options, for resources are always limited irrespective of a country's fiscal circumstance, and effect size and the factors that may make an intervention important or not are two critical scientific considerations that need to be considered.

If dollars are to be spent on an intervention, the trade-off is that dollars are not available for another intervention for the same or a different outcome – that is the opportunity cost and is the basic reality of policy formation. Yet political advocacy played out in the media and elsewhere often obscures this reality.

We have seen this debate come to the fore in recent discussions over class size in schools versus other interventions. No one argues over the desirability of improving educational outcomes even in times of relative fiscal constraint, but some think that there is a need to keep class sizes small for better outcomes and others suggest that the effect of smaller class size is trivial compared to other ways of enhancing pupil outcomes. This is a question that is readily amenable to empirical research, but other factors affect the decision-making process. A further issue that can arise in such research is that it should be subject to peer review before being incorporated into the policy process – there is a need to separate the undertaking of the research from its evaluation for the purpose of policy formation. I shall return to this matter in a subsequent report.

The related issue here is to be clear about what outcomes of education are meaningful – is it student happiness, formal school performance or should we be looking at employment potential or progress through subsequent educational experience or should we even be looking long-term at variables related to integration into society (such as employment history, earnings or stable relationships)? The problem is that the last few measures may be what really matter, and that examination performance may be at best a surrogate measure of uncertain quality for predicting societal success. This example highlights the value of a more sophisticated and informed discussion that will improve outcomes for children and for society.

Box 4: How to target investment in young children?

Another contemporary example of context is that of intensive early childhood education of the type that involves both home and institutional activities, which is very expensive. There is good evidence that intensive early childhood interventions are very effective in terms of long-term outcomes such as measures of employment and greater social success (for example, fewer arrests) for those of greatest vulnerability but have very limited if any long-term benefit for those of low social vulnerability. There are, however, other general purposes of early childhood education, such as parental relief and short-term effects on school readiness and performance. Here then is the policy dilemma. Should a government have a plan for early childhood education which is designed around equality of access and provision, which means that the general purposes shape policy, or should a government focus on using early childhood education as a tool for reducing longer-term morbidity and place a larger component of funding on a smaller group of children to enable equity of long-term outcomes? Looked at from an analytical perspective, the latter would make sense from the perspective of human capital development, but the other advantages of early childhood education have to be considered. The choice is thus a values-laden one expressed through the political process, and how those options are resolved will depend, in part, on how the evidence is presented to the public and politician.

9. When is the science settled?

Should we ever assert that 'the science is settled'? In a strictly formal sense, certainly not. Science is a process of 'organised scepticism' by which working hypotheses are subject to refinement or replacement as new evidence comes to light – the 'approach the truth by successive approximations' of Bertrand Russell or the 'scientific revolutions' of Thomas Kuhn – and in a critical distinction from dogma, absolute scientific truth is rarely if ever proclaimed. That does not mean however that strong scientific conclusions can never be reached. Indeed the processes this paper refers to allow strong conclusions to be formed even in the most complex areas.

Science is, however, always open to modification by new observations. The role of scientific advice is to opine when a particular position has been established with sufficient confidence to allow practical application to proceed. A simple example might be Newton's laws of motion, which for 200 years provided a solid theoretical basis for explaining many physical phenomena and underlaid many of the mechanical inventions underlying the Industrial Revolution; but 100 years ago, Einstein with his theory of special relativity showed that Newton's laws are only an approximation valid for the speeds and masses that are familiar to us. Corrections for relativistic effects are now finding practical application, for example in the design of satellite navigation systems.

Governments must often act on incomplete knowledge, since issues of risk and uncertainty arise in dealing with complex physical, biological, social and economic systems and science cannot provide certainty in dealing with such policy-related questions. Recent and familiar examples include how should the effects of climate change be factored into long-term planning of infrastructure and how to deal with incursions of agricultural pests. Modelling using approximations to the real world allows scientists to test the bounds of probability to understand what could happen and what are the likelihoods of the outcomes when systems (such as that of our climate) are complex. Models are by their nature only imperfect representations of the real world and a margin of error is therefore expected – even so they can provide key insight as to the boundaries for future predictions if used correctly.

Where interventions are based on such evidence it is vital that on-going assessment is made to check for changes outside of original predictions or knowledge (such as new technologies appearing, dramatic changes in prices of resources, or new feedbacks or effects being discovered). Further where decisions are taken in a setting of uncertainty, it is critical to evaluate in an on-going matter whether the chosen intervention will work or not. This should not be taken or used as a criticism of a particular policy decision or of the underlying science, it simply reflects the practicalities of the limitations of knowledge. Indeed such uncertainty is almost inevitable and requires greater and pre-emptive consideration of evaluation when new programmes are introduced.

10. Spill-over effects

In any intervention, there is also the potential for spill-over or 'side' effects which may be positive or negative – and the way these are assessed also requires careful approaches. If the study is not designed to look for these they may remain unknown. The issue of side effects is critical to the policy maker – they need to consider what are the good and bad side effects of any intervention and avoid 'unintended consequences'. For example, if a policy intervention to reduce carbon emissions by constructing more energy-efficient homes leads to consumers saving fuel and therefore money, will they spend the saved money on a holiday abroad which itself has an associated environmental impact? Again there are ways to design interventional monitoring to look for side effects and assess how important they are.

11. Proving the negative

Opponents of the introduction of some new technology often demand absolute proof of its safety. As discussed above, some philosophers of science have argued that this is not formally possible, since 'proof' of a scientific fact can be no more than the current consensus on the interpretation of the existing set of evidence. Similarly, proving that any technology has no adverse impacts can never be proven to be true, but can only be proven not to be true. Observations can show no adverse impacts, but this does not rule out that some other set of observations in the future may show some such effect. Society is often poorly positioned to balance the prospect of a small immediate risk versus large future benefits. The only rational approach must be one of risk and hazard assessment and adjudicating on whether the technology can be managed appropriately or not. Otherwise no new technology would ever be introduced.

Box 5: Shifting the paradigm

Up to the 1980s, the consensus of the medical profession was that peptic ulcers were caused by excess acid in the stomach, perhaps brought on by stress or spicy foods. Treatment consisted of acid-suppressing drugs, and any idea that ulcers could have an infectious origin was dismissed because bacteria were thought to be unable to live in the acid environment of the stomach. Then, Australians Barry Marshall and Robin Warren discovered the involvement of *Helicobacter pylori*, an acid-tolerating bacterium able to colonise the stomach lining, in gastric inflammation and ulcer formation (one of their experiments involved Marshall deliberately infecting himself with *Helicobacter*). Although acceptance of the novel findings was slow, the role of infection was eventually accepted, Marshall and Warren shared the Nobel Prize in Medicine or Physiology, and antibiotics are now first-line treatment for peptic ulcer.

The philosopher of science Karl Popper[1] famously gave the example of the statement 'all swans are white' to illustrate this point – to prove this statement would require laborious examination of the world's entire population of swans, but to disprove the statement requires only discovery of a single black swan. We might not wish to forego the benefits of a new technology until, to continue the metaphor, all swans have been examined, but at the same time we should be alert to the appearance of a black swan.

There are particular issues around how to consider the risks of very rare but high-impact events (to which the term 'black swan' has also been applied). Often we are more concerned about the possibility of a very rare event which has high impact when in reality more common events of familiarity have greater risk – statistically, driving a car is much more dangerous than air travel, but people worry much more about air crashes than they do about their daily car journey.

In a subsequent discussion paper I will consider these issues in more detail.

12. From the laboratory to the real world

Another issue to consider is whether the effect shown is applicable to the real world. Is a psychological study in rats applicable to humans? Can the beneficial effects of a medicine shown in a carefully controlled clinical trial be extrapolated to the real world where patients forget to take their tablets and variability is far greater than in the trial cohort? Will an alternative energy generation system really deliver usable power when connected to the grid? Will a social intervention piloted in the US be applicable to New Zealand? All too often the answer is no or not certain, underlining the importance of knowing the context of a particular study and of distinguishing between *efficacy* (whether a particular technology works under ideal conditions) and *effectiveness* (whether it delivers its benefits to the target population or situation). Here scientific expertise is necessary to assist the policy maker.

A related issue is one of 'going to scale'. Often an intervention may work in a pilot situation but when applied more generally it fails. This is understandable, as pilot projects often involve enthusiastic advocates whose attention to detail cannot be generalised. This is often not appreciated by those advocates and is a reason why all programmes after piloting should continue to be monitored to evaluate whether effectiveness is maintained. Such monitoring can often identify what are the critical success factors.

13. Plausibility and track record

Modern science is a team activity that proceeds by building on and testing previous work – as Isaac Newton put it, "standing on the shoulders of giants". In modern medical, physical and biological sciences at least, the 'lone maverick' with observations that challenge a whole canon of previous work is an unlikely – although not unprecedented (see Box 5) – pathway to valid new knowledge. So, a factor to take into account when assessing a study is its plausibility – how well does it fit with what is already known about the system under study – and the track record of the research team that produced it – do they have a record of achievement in this or a closely related area? Again expert analysis and replication are core to validation.

14. Conflicts of interest

A prerequisite for the credibility of scientific knowledge is that it is obtained and analysed in an unbiased manner. Bias can derive from personal, political, financial or value-based influences. If researchers are perceived as subject to potential biases, then there is a risk that their objectivity in presenting or discussing their research

[1] I am well aware that some philosophers of science question the value of Popper's deductive or 'falsification' approach, but it serves here to illustrate the point.

results may be seen as compromised. Many research fields – particularly in the biomedical sciences – have developed procedures to try and ensure that conflicts of interest are identified and managed so that appropriate conclusions about a study's objectivity may be drawn. Potential conflicts of interest may arise, for example, when researchers have investment or employment interests with commercial entities that stand to benefit financially from the results of the research, when they are dedicated members of a particular lobby group, or where a study is sponsored by a values-based organisation such as an environmental pressure group.

The existence of such relationships does not necessarily compromise the study or the scientist. Nevertheless, it is important that they are revealed through statements of authorship or financial support so that readers can judge the credibility of the work.

15. Scientific error, fraud and publication bias

Science is a human endeavour and both accidental and intentional bad practice does occur.

Scientific fraud occurs when scientists misrepresent the collection or analysis of their data. This can range from outright fabrication of results, through biased data analysis (for example omission of data that do not fit the hypothesis), to plagiarism (using the ideas or words of other scientists without acknowledgement). Although the processes of science are designed to minimise the risk of fraud, it can never be completely obviated. The peer review process is critical in protection against the problems of error and fraud and thus there is rightfully suspicion when scientific claims are made in advance of peer review. But peer review is not perfect and errors, intentional or accidental, do appear in the literature. Replication is the key protection and for these reasons when a surprising result is found, the scientific community needs to remain sceptical until there is independent replication. Unfortunately, there is evidence that scientific fraud is more common than previously assumed, with the changing nature of the scientific endeavour and of the scientific career putting pressure on scientists to achieve. Even without deliberate fraud, such pressure can cause scientists to make premature claims for their work, with increasing use of 'publication by press release' that tends to exaggerate what has been found.

There can be bias in scientific publication: replications and negative results are usually harder to publish than initial observations and positive results. Sometimes conflicts of interest (see above) can lead to non-publication of work potentially damaging to a researcher's pet hypothesis or study sponsor. For these reasons, there has been a recent move in medical research to register clinical trials prospectively, placing all trials in the public domain and encouraging prompt disclosure of results.

16. Science and values

As noted above, the process by which scientific knowledge is obtained should be as free as possible from value-based influences.

Difficulties can occur when scientific knowledge confronts values in any debate about the policy agenda. We often see weak studies being used to promote a particular point of view, and unsubstantiated anecdote from a few vocal individuals or advocates being given the same weight as a body of carefully performed and reviewed research that may not have been presented in an accessible way. The tendency for the modern media to want to create and magnify controversy rather than be transmitters of information is a further problem. This can be manifested in the 'false Wbalance' of giving equal time to maverick or proxy claims and to those who represent the state of scientific consensus. This in fact is biasing the debate against the consensual position and does the public a mis-service.

Issues are confounded even more when discussion about complex science becomes a proxy for debate that is re-

Box 6: Beyond received wisdom – a taxing problem

For the first part of the 20th century, labour legislation and consequently tax policy assumed that families in the main had one earner. This received wisdom determined the form of the tax allowances that reduced the tax liability of low-income households. From the early 1960s, an increasing share of the paid work force consisted of women as second earners, usually working part-time. In 1977, following the first results of a comprehensive income questionnaire attached to the annual Household Expenditure Survey, the sources of income of households were analysed. The research showed that the received wisdom implicit in the old tax allowances was quite wrong, and that most of the lowest income households had two low-income earners, rendering them ineligible for the tax relief that was intended to be targeted at them. In the next budget, tax allowances were changed to focus on the total family income.

ally not about the science but is a debate about values – we see that in issues such as climate change and the regulation of 'natural' health products. Similarly, while there are appropriate but diverse values-based opinions regarding genetic modification of food, attitudes to it have been largely shaped by exaggerated and now invalidated claims of health concerns. Even so, the false science continues to be presented as the basis of rejection rather than a genuine discussion being held on appropriate grounds (philosophical, economic or ecological).

Such debates can be compounded by poor scientific literacy that can aggravate the risks of mis-communication, intentionally or accidentally. The need to increase scientific literacy is key to a society making their choices about both new and old technologies or incorporating new knowledge into public policy. The scientific community has an obligation to engage more proactively with the community, particularly in ensuring an understanding of new technologies early in their emergence.

17. The need for science and research applied to public policy

These issues become more critical in considering interventional programmes in areas such as social welfare and education (see Box 6). The nature of the public policy process is such that many programmes are introduced without consolidated and validated evidence and on the basis of anecdote, belief or bias. These subjective factors may well provide a valid basis for a trial to see if the hypothesis is supported, but they do not guarantee that the desired effect will be achieved. For example, driver education was introduced into high school programmes in the belief it would reduce teenage road accidents, but in fact it did not and actually increased accident rates. It seems logical to introduce drug and alcohol education into schools, but some formats of such education have been shown to increase rather than reduce drug usage. These simple examples highlight the need for scientific approaches to be applied to much more of public policy when new programmes are introduced. It is important that such research is of quality and that its limitations are understood. There is an increasing understanding of the need to think about the issue of evaluation before programmes are introduced – this may influence the baseline data that needs to be collected and the mode of introduction of a new programme. The public interest is best served by research that demonstrates effectiveness rather than by assuming effectiveness based on opinion and anecdote.

18. Conclusions

Many decisions are made at various levels in our society, from fluoridation of the water supply to dealing with climate change, where the underlying scientific data can be well used or misused. Good decision-making requires that both the public and policy makers are informed as to the quality of the evidence. This requires a media that understands its responsibility and more effective communication between the scientific community and the general public. Only then can evidence form the foundations on which other value-laden considerations such as fiscal and ideological factors are overlaid – this is properly part of decision making in a democracy. Public opinion is central to policy formation in a participatory democracy: that is why the public requires an understanding of how data can be well used or misused, how advocacy can create confusion, intentionally or otherwise, and why it is that science can appear to be used or misused by both sides of a contentious argument.

The better application of science in policy formation should be as free as possible from that advocacy component. It should inform about what we know and do not know, what is effective and what is not, leaving the values domains to the public, officials and politicians. If science is well done and properly analysed in context, then it is not science that is the problem. Rather it is more often how that science is used or misused and presented. Science has its limits – it cannot always give precision, and even when it is well done it will in many cases be dealing with probabilities and uncertainties, but these limitations are not an excuse to ignore or misuse science.

Acknowledgements

I thank Dr Stephen Goldson FRSNZ, Dr Alan Beedle and Ms Kate Harland from my office for their contribution to developing this paper. Professor Sir David Skegg FRSNZ (Royal Society of New Zealand), Mr Len Cook (Families Commission) and Ms Jacquie Bay (Liggins Educational Network for Science, Liggins Institute, Auckland) peer reviewed the paper and made numerous helpful suggestions.