For pluralism in scientific method

John C. Ashton
Department of Pharmacology & Toxicology, Otago School of Medical Sciences, University of Otago, Dunedin

Even since Galileo, philosophers have tried to explain why science has been so successful. All of the great philosophers in history have taken an interest in epistemology – the philosophy of knowledge – and the most important philosophers of recent centuries have grappled with the problem of scientific knowledge. At the same time, probability theorists and mathematicians have also grappled with the problem of scientific experimentation, giving rise to modern statistical methods of experimentation and analysis. Is there then a 'right' way to do science, or is science a matter of 'anything goes' as Paul Feyerabend has argued (Feyerabend 2010). Feyerabend was writing in reaction to the rigidity of the methodology of Imre Lakatos, who in turn was reacting to the writings of his teacher Karl Popper. But this context has been lost in some recent popular histories of science, which have presented science as a wholly anarchic enterprise (Brooks 2012). It is my contention that there is not a *single* right way of doing science, but nor does 'anything go'. Rather, a plurality of approaches to science is possible.

Classical science and normal science

The focus of this article will be on only two approaches to science. The first I will call classical science, and the second could be called 'normal' science. The latter term comes from Thomas Kuhn's description of science (Kuhn 1962), and I use it to include operational and applied research, industrial research, and any kind of puzzle-solving research carried out on a big industrial model. By classical science I mean science as the quest to discover explanations of the way the universe we live in works, as typified by the great works of the scientists of the Enlightenment. The distinction is not perfect; much of modern astronomy and cosmology is both classical science and 'big science'. Nevertheless I will argue that the distinction is real and useful. Indeed it has been described in terms of a battle for the very 'soul of science' (Fuller 2003), though I will argue that both approaches have particular strengths for particular aims. In particular I will argue that many debates over methodology in science and in statistical analysis can be resolved by reference to these distinctions.

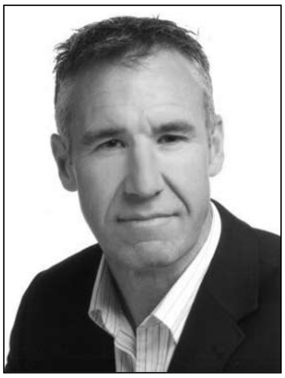
Classical science involves the close interaction of theory and experiment, the aim being to uncover invisible worlds behind the world of appearances; it is about explaining the seen in terms of the unseen. The hidden worlds that have been uncovered by classical science include the worlds of the atom, the cell, deep evolutionary and geological time, biochemical structures, the structure of the neuron, deep space, the interior of the earth, and the shifting plates of the earth's crust, to name but a few. Classical science is therefore disciplined and constrained in its speculations, not only by experiment but also by the search for good explanations[1]. I contrast this with a more modern style of science modelled on operational research – i.e. research for industry, production, design, engineering, or some other practical application. Accompanying the operational research model are the statistical methods that are used to increase efficiency in the testing of models and procedures. I argue that these methods, though powerful for their given purpose, can sometimes inadvertently be a hindrance to scientific discovery in the classical mode. This is because the perceived authority of such methods can tempt researchers into carrying out 'explanationless science', as discussed below.

Experiments test explanatory theories in classical science

A continuing source of ambiguity in discussions over scientific method is over the use of the term 'hypothesis'. In classical science a researcher has some phenomena to explain and posits a theory to do this. Predictions are drawn from the theory, and when these are tested experimentally they are called hypotheses. Therefore, experiments never directly engage with a theory, but only indirectly through the intermediary of predictions/hypotheses. By contrast, in 'explanationless science' the hypothesis *is* the theory, often consisting of little more than conjectures

** Correspondence: john.ashton@otago.ac.nz*

[1] See David Deutsch's (2011) book 'The Beginning of Infinity' for a spirited



John Ashton is a senior lecturer and researcher in pharmacology at the Otago School of Medical Sciences, where he first started work as a research fellow in 2001. He is currently researching cannabinoid drugs and the endocannabinoid system, the pharmacology of pain, and drug interventions into stroke. Dr Ashton has worked in other areas of experimental sciences, such as stem cell research and zoology.

about simple associations between groups of variables. In this approach the associations measured in an experiment are synonymous with the theory under investigation. It is interesting to trace the history of such a notion back to empiricism and associationism in 17th and 18th Century philosophy. A detailed discussion of this is beyond the scope of this article, but it is important in understanding the crucial role of the informative *content* of theories in science.

Hypothesis testing in operational research in comparison to explanatory science

The modern statistical experiment probably began with R.A. Fisher, who provided a solution for the problem of controlling for random variables in experimentation, using an integrated set of procedures governing experimental design, data collection, and data analysis (Marks 2003). Fisher designed his approach to experimentation to solve a problem in operational research; in his case agricultural research. Although as a theoretical geneticist Fisher provided the world with theories of impressive explanatory power, in his role as an agricultural researcher he tested relatively simple 'theories' consisting of no more (for example) than the association between variables such as fertiliser and plant growth. Therefore, his agricultural experiments directly tested a simple idea of practical interest, namely how much some factor could contribute to plant production. This is in contrast to experiments designed to test explanatory theories. To use a much discussed example, Eddington's experiment measured the effects of the sun's gravitation influence on the path taken by light from stars on the way to earth. But this in itself was not of primary interest. What mattered was that the experiment tested a prediction derived from Einstein's General Theory of Relativity, the latter with its capacity to explain a vast richness of natural phenomena. Many similar examples could be given, but the key difference is that in the agricultural experiment the variance in the data across the groups defines the probability space (the range of possible outcomes over which probabilities are calculated) whereas Eddington's experiment – though simple in terms of variation in the data collected – pertains to a much wider probability space defined by the theory that it is designed to test. That is to say, the precise direction and magnitude of the deviation of the light path with respect to the space and time it was measured as predicted from Einstein's theory is highly improbable *a priori*, given the enormous range of alternative light paths that could conceivably be measured. Therefore, a *P* value generated for Eddington's data by comparing one set of measurements with (say) control readings radically underrepresents the improbability of the result. This crucial element in the interpretation of experiments is lacking from a *purely* statistical model of science (i.e. explanationless science).

Considered as an idealised statistical experiment, Eddington's experiment would take this form: control readings for the apparent position of a star when its light path does not pass close to the sun are gathered. Then, at the time of an eclipse, measurements are taken as light from the star passes close to the sun. The apparent position of the star is calculated for the two groups and compared with a simple statistical test and a *P* value is generated. Of course, how well measurements of the star's apparent position matches predicted values could be calculated using an even simpler test, but the control readings are necessary as a check on whether the apparatus is adequate to the job of measuring the apparent position of stars. But any *P* values calculated only take into account the variation within and between sets of measurements, with or without considering their deviation from values predicted from theory. What is missing from the account is that the apparent shift in position of the star due to the gravitational influence of the sun is predicted with great *precision*. Considered as polar coordinates relative to the observer, Einstein's theory predicts a precise direction and magnitude for the apparent shift. This selects out of a set of possibilities that include (at least) the entire field of view. The variation of the experimental readings only covers a tiny part of the space of possible measurements, and no consideration of effect sizes can substitute for the extreme improbability of the result of the experiment considered with respect to all conceivable measurements[2].

In classical science, powerful experiments come from powerful theories. The meaning and probability of experimental results transcend anything calculable from variation in the data, and can only be approached conceptually by considering the specificity of the prediction considered as a subset of all conceivable outcomes. This goes beyond the idea of 'effect size', a concept that applies to scalar experiments irrespective of theory. But, as operational experiments are often not tests of predictions from theories, and as they test (almost) content-less hypotheses, they generate *P* values that come closer to a full description of the probability of the result, and so can be used in rational decision making. What defines a good experiment in classical science is not so completely captured by statistical design principles, but requires careful consideration of the logical structure of the theory under critical consideration.

In operational research, this lack of accounting for the explanatory power of theories is not a problem, and in fact the statistical model is ideal because what is under investigation is the ability of some procedure to generate a desired result: *the output.* But in classical scientific experiments, the outcome measure is only indirectly related to the content of the theory. The purpose of a scientific theory is to explain data. The purpose of operational research is to generate desired outputs: the goal being procedures, devices, machines or mechanisms, models of various description that are good at a particular function. Operational research produces things that are well designed for a task, in much the same way that biological adaptations are well constructed to carry out a particular task. It follows from this that the methods of science must deal with the explanatory power of scientific theories whereas those of operational research do not. One of the key discoveries of Karl Popper was that the explanatory power of a theory is in inverse proportion to its logical probability – the very property of Einstein's predictions that are discussed above. Fisherian hypothesis testing *in isolation* does

[2] *This argument has a Bayesian flavour, but it is incorrect to think that Bayes' theorem can be therefore used to calculate a probability for the truth of Einstein's theory. A theory with good explanatory power will have greater content and lower probability than any of its predictions, so that the prior probability crushingly discounts any evidential posterior probability. But this does not matter, as what is really at stake is that the theory has passed an extremely severe critical test. It makes as much sense to say that Einstein's theory is probably true as it does to say that the Bugatti Veyron is probably the fastest production car that will ever be built. All that can be said is that this car is probably faster than anything else yet built, and that considered as a random ensemble of matter, it is improbably good at hurtling across tarmac. Similar language can be used for the explanatory power of Einstein's theory.*

not deal with explanatory power; in fact, its procedures deal with hypotheses that are all but stripped of explanatory power – theory-less or explanationless hypotheses that describe but do not explain simple associations. What it does do very well is help in making decisions on experiments designed to test the functioning of some operational model.

Classical science is about discovery and explanation whereas operational research is often about optimisation

A theory with high explanatory power rules out a great many possible outcomes, whereas a theory with low explanatory power rules out relatively few. This quality of theories is important in interpreting experiments that are designed to test them. Theories with high explanatory power, should they pass the experimental test, warrant our profound attention. But when there are many possible ways that an experimental result could have been true regardless of the theory, then the passing of an experimental test counts for little, because the likelihood of it surviving the test, even if false, is high anyway. It is possible to conceive of experiments that test two theories that differ in this way, but that generate identical *P* values, effect sizes, and other statistics. Hence the ascendance of the statistical model of experimentation can have dangers for classical science if it lures scientists into explanationless research, investigating mere correlations between variables divorced from explanatory context. When explanatory power doesn't matter, explanations won't be sought. When logically improbable theories are not valued, discovery will be derailed. If scientists strip their theories' predictions of content, shoe-horning them into a form easily amenable to a pre-packaged statistical model, then it is possible to denude the experiment of its interpretive context. Because statistical experimental design has all the trappings of epistemological strength, the statistical model may be used to achieve some kind of apparent 'experimental power' in the absence of any explanatory theory to be tested. This encourages empty, virtually theory-less experimentation, where the sophistication of mathematical design becomes a substitute for scientific reasoning and a smokescreen for the lack of any significant theoretical content.

In the type of science typified by operational research this is not a problem; given that its goals are often good design rather than good explanation. The constructs and procedures under investigation have little or no explanatory content. The distinction becomes even more important for multivariate statistics. In classical science, mathematical theories with large numbers of parameters are often frowned upon. Such theories are very hard to disprove if wrong, as changing the parameter constants can be used to fit the model to nearly any data. Theories become more complicated than the data to be explained, and so explain nothing. The complexity of such models make them opaque to critical reason, providing less insight into the hidden processes of nature than the data to which they pertain.

However, what is a weakness in classical science can be a strength in operational research, where empirical feedback is used to correct a complicated model in a way that reason cannot do. When explanation is not the goal, a highly complex structure (mathematical or physical or procedural, etc.) is often favoured because it can be adjusted in many ways to produce better results (for example, a fit to meteorological data) without having to be replaced wholesale. The experiment tests the operational procedure or model to be optimised, with each experiment a decision making unit for further optimisations of the model. The model may not be falsified and replaced (as theories are in science) but be adjusted after each experiment to better carry out its task. By contrast, and as argued above, this can be a disaster for classical science and theoretical understanding[3].

A case study: Two ways of using *P* values in science

In order to make the distinction between the two types of science concrete, I will use as a reference point an on-going debate within science – the use of *P* values and null hypothesis testing. Seemingly a minor issue, in fact the topic rouses passions to the extent that some leading journals (e.g. *Epidemiology*) refuse to publish *P* values. I will argue here that the argument is not really about *P* values, but about how science should be done.

For many decades, statisticians have expressed concern, even bewilderment, over the extensive use and perceived misuse of null hypothesis testing in science, and the generation of *P* values, at the expense of more informative analyses[4]. Statisticians fear that scientists' misinterpretation of hypothesis testing gives an illusion of objectivity. However, although there is much truth in this, the extensive use of *P* values wouldn't have continued under such sustained critical attack unless they carried out some useful function in science. Sinclair, who provided a perceptive explanation for scientists' fondness for *P* values that has been largely overlooked (Sinclair 1988), argued that scientists use *P* values in a different way from that intended by statistical theory. Formally, *P* values are instruments for statistical decision making, with conventional thresholds for *P* values used as criteria for decisions as to whether to reject a null hypothesis. Sinclair pointed out that this is not how scientists often use *P* values; they are instead used as a sliding scale to flag instances where a signal may be tentatively judged to have been detected among the noise. To paraphrase Sinclair, these values are not used to make any objective decisions about null hypotheses or about any particular experiment, but are rather descriptive terms that are gathered as clues in a process of scientific inference that takes place over the course of papers, research programmes, and critical debates in the literature. Scientists, in other words, need not make inferences from experiments like statisticians do, but make tentative decisions – that take into account all evidence that is at hand – widely understood.

Understood this way, *P* values are useful descriptions of data, as they are independent of degrees of freedom, summarising information in a single dimensionless number between zero and one. Therefore. one function of *P* values as used by some scientists is not for comparison to conventional thresholds for decision making, but for communication. Sinclair pointed out that when *P* values are used in this way, then it makes little sense to correct for multiple comparisons within an experiment. Standard practice is to perform a correction on *P* values for

[3] *Not all quantitative methods are blind to explanatory power. For example, in information theoretic methods for data analysis, increasing numbers of parameters in a model discounts its informative content.*

[4] *A comprehensive list of quotes and references is at http://www.indiana.edu/~stigtsts/quotsagn.html (accessed 27 June 2012).*

multiple comparisons (post-tests) such that the probability of finding at least one comparison in the experiment at a certain *P* value is made to equal that nominal value. Sinclair pointed out that this assumes that the experiment is a unit over which type I error should be normalised. But this is merely an assumption of statistical decision making, and makes no sense for how some scientists view their data, comparing and cross-checking within experiments, across experiments, and across papers and even disciplines. Unlike in statistical theory, in classical science there is no single experimental unit for decision making, and no natural unit over which type I error should be normalised.

Experiments are decision making units in statistical theory but not necessarily so in classical science

Histories of science are replete with dramatic experiments that 'changed the course of science'. Although such experiments make good history, they are not necessarily representative of experimentation in science as a whole, where experimentation is a much messier affair, involving a tinkering process of problem solving by trial and error. Indeed, the image of history-changing experiments may be a literary fiction, used to summarise a much longer and less balanced series of trials and error. The theory that singular experiments (rather than series of experiments in the sense of trial and error) should be used as decision making units also probably has its origins again with R.A. Fisher, as discussed above (Marks 2003). The Fisherian experimental procedure involves randomisation of subjects to groups such that only data collected within one experiment can be analysed for statistical error using methods devised by Fisher specifically for the task. Therefore, when *P* values are used for statistical hypothesis testing, experiments are the units over which decisions should be made. Often, an experiment using this approach is a test of something that will be used in a practical application. Therefore, *P* values would be normalised across each experiment so that they can be used to make decisions and to quantify risk of failure when used in calculations in other steps in the design process, production, marketing, or some other aspect of application.

However, classical scientists are primarily interested in explanations, and, freed from having to make digital decisions on an experiment-by-experiment basis, they look for contrasts and comparisons in data at all levels: within experiments, between experiments, or across research papers. In this way a richer description of the problems that a successful explanation must solve is gained than anything that can be provided by a single experiment. *P* values are useful tools for communication in this process. The statistical model of experimentation common to much of normal science is responsible for a view of experimental science that misses the crucial role of the interaction of the results of one experimenter's work with another. Decisions are not made by reference to experiments or papers in classical science, no matter how tiny reported *P* values may be.

The case study continued: Effect sizes and *P* values

Another way in which scientists are often criticised in the way they use statistical methods is in reporting *P* values without reporting any corresponding effect size (such as coefficients of determination, or standardised differences between means, and so on). There is a lot of truth in this criticism, but taken too far it misses an essential point and betrays a standpoint that comes from explanationless science. That is, effect sizes, and strengths of associations between variables are not meaningful in *classical science* in themselves, but only with respect to the explanatory theories to which they pertain. A very small effect or loose association may still be very important when seen in the light of a particular theory. It is only in operational research, where the goal is often maximisation or optimisation of some output measure, that effect sizes are interesting in isolation.

Because the meaning of an 'effect' in classical science depends on the explanatory theory at stake, a case can be made that what is actually most important is whether the detection of the effect, at any scale, is due to chance. Therefore, *P* values take on a particularly important *critical* role, flagging instances where researchers may be in danger of fooling themselves when mere chance is at work. In other words, *P* values carry out a *decision making* function in operational research, but a *critical* function in classical science.

The central importance of explanations in classical science, and the role they play in disciplining and constraining interpretations of experimental results can also help to resolve a central paradox of the statistical model of science. The paradox can be expressed several ways: first, the problem of 'over-fitting', where repeated attempts at statistical model fitting is almost bound to produce random 'fits' if continued long enough with enough variability in the model parameters; second, with any threshold value for statistical significance, repeated null hypothesis testing will generate many false positive results (type I error). The paradox comes when attempts are made to overcome these sources of error by various conventions for restricting the range of comparisons and models tested that the researcher is permitted to make. Hence, a commonly heard piece of advice from statisticians is to choose a set of comparisons or models before the analysis, and stick to them. But reflection shows that this is arbitrary – as if somehow the temporal order of steps in a purely logical process could have any bearing on the truth of things. Indeed, the advice comes very close to subjectivism. There is no final way to normalise error rates, provide consistent rules to prevent researchers from 'fishing' for correlations or over-fit models. All these things are not an abuse of statistics, but a limitation of a particular approach to science. The answer to the problem lies in recognising that the problem is inherent in explanationless science, and that when experimental analyses are constrained by the requirement to provide *good explanations*, then many of the problems resolve.

The use of conventions to limit the number of comparisons that are made in statistical analysis also seems to be a feature of confirmationism. That is, an assumption seems to be that, if few comparisons are made, and yet a small *P* value is found, this is very likely to be due to more than coincidence and so the hypothesis is thought to be confirmed. But, this is a chimera, because, given a large research community and the very iterative hypothesis testing process that defines the discipline, coincidences will repeatedly occur. By contrast, a falsificationist approach to science looks to criticise hypotheses and explanations, so that *lack of* statistical significance becomes the main point of interest. Falsificationist science is intimately related to classical science, with its emphasis on explanations. In classical science, statistical analysis is carried out in the context of providing critical tests for explanatory theories – not simply correlative hypotheses – and to test whether predicted relationships are as great as that expected from theory (and whether apparent relationships are due to chance). Because good explanations have certain properties such that they are not endlessly variable (Popper 1959; Deutsch 2011), the paradoxes of statistics can be resolved in classical science. Similarly, cautions about correlations not proving causation is another bogey of explanationless science. Causation is a feature of explanations, never directly measured in any experiment regardless of the quality of experimental design and controls. A classic example of the clash between the two approaches to science can be seen from the debate between R.A. Fisher and those who followed the reasoning of A. Bradford Hill over the dangers of smoking (*see* Le Fanu 2002; McGrayne 2012). Fisher maintained that only correlation not causation had been shown, and that there was no evidence for harmful consequences from smoking. But Bradford Hill had used logical and theoretical reasoning in his interpretations of data, going beyond the statistical model used by Fisher. On this issue Bradford Hill was right and Fisher was wrong.

If *P* values are seen for what they are – decision making tools in explanationless operational science or critical tools in classical science, not numbers that exhaustively capture the uncertainty inherent in an experiment – they become just one useful tool in the researchers toolkit. Where problems arise is when the notion is entertained that *P* values take into account the entire set of possibilities that are relevant when evaluating an experiment, a feature of explanationless science. This is an abuse of null hypothesis testing when it is done in classical science, where experiments are always evaluated in the context of theories, and where the *logical* probability of predictions must be carefully considered (see above).

Classical science, normal science, induction, and falsification

So, in a falsificationist philosophy of science, statistics can only be used to criticise but not confirm theories. In operational research, where the truth of theories (only the performance of models) is not at stake, this issue doesn't arise, but it is fundamental to the process of explanation and discovery. To see how this affects science, consider as an example the use of statin drugs to treat cardiovascular disease (CVD). Statins were first hypothesised to reduce the risk of CVD on the basis of the hyperlipidemia theory of CVD, which had earlier been proposed on the basis of fortuitous discoveries. Statins have been very effective in large-scale randomised clinical trials (RCTs) (Maggo *et al*. 2012). As operational research, this is all that need be said: statins work. But from another perspective the results of the RCTs look quite different. Have the statins passed the test? Yes. Does this mean that the hyperlipidemia theory of CVD has been confirmed? No. From another perspective it is the *criticisms* of the theory that have been uncovered by the statin RCTs that are most interesting: for example, the repudiation of the idea that statins are only helpful to people with elevated blood lipids. Therefore, although the RCTs validated statins as medicines, from a scientific perspective they remain problematic. As classical scientific experiments, the statin RCTs have performed a *critical* function, exposing gaps in understanding that could lead to deeper more comprehensive theories about the origins, nature, and treatment of CVD.

It is important to acknowledge here that one of the great triumphs of statistical experimentation is the randomised clinical trial (RCT) in medicine. These are exercises in operational research *par excellence*, where hypotheses describe a procedure to be optimised; the procedure being the administration of some treatment and the output being some measure of health or disease. There need not be any explanatory theory to which the test medicine pertains; much drug research is, in the jargon of medical researchers, 'purely empirical'. On the other hand, RCTs are not good mechanisms for drug discovery, which tend to happen more by fundamental research into the explanations behind pathologies, tests of bold theories about new treatments, or simply by accident, things for which there is often little room in the strict confines of large-scale experimentation. Where RCTs have been extremely useful for classical science is in the *falsification* of hypotheses about medicines, dismissing traditional treatments that are found to be ineffective. Similarly, RCTs can provide powerful criticisms of explanatory biomedical theories (such as in the hyperlipidemia theory discussed above). Therefore, the full power of statistical experimentation is realised when the analyses are closely linked to explanatory theories, and used in attempts to *criticise* such theories.

Scientists deal with questions and their answers, not with decisions

Science is a problem-solving process; conjectures are made to try to solve a problem, tested, and refuted, and new conjectures are made. So much is now widely accepted. However, this view of experimentation, though close to the truth, is not quite correct. What happens more often than not in a scientist's working day is that an experiment returns not (or not only) an answer to a question, but another question (or many questions). Karl Popper is, of course, well known for his discussion of the logical asymmetry that explains why the arrow of logic in experimentation runs in the reverse direction to that which was once supposed, i.e. data are not used to induce theories. Rather, theories are proposed and only falsified (but never confirmed) by data. Less well known is another inversion, a theme of much of Popper's later writings. Popper maintained that the actual course of science is that it starts with problems or questions, runs through conjectures and tests at each step, proceeding to more questions and problems (e.g. Popper 1963). This seems to me to be a profoundly correct description of experimental science, and does much to explain exactly how scientists go about the business of solving problems, making discoveries and explaining things. Indeed, perhaps public trust in science may be increased if it ceases to be identified with the results of experimental 'findings' but is understood in terms of the error-correcting problem-solving discovery process that it is.

However, in many of the descriptions of the hypothetico-deductive method, a picture is presented of iterative hypothesis testing, where conjectures are presented, refuted, and replaced with a new conjecture. This view of problem solving has caused much hand-wringing by many of Popper's followers, who have agonised over whether or not it is correct to call one conjecture closer to the truth than its predecessor. But it was not a large issue for Popper himself, as Popper saw science as an evolving set of questions as much as of theories. The step-by-step elimination of hypotheses experiment-by-experiment may fit well with the statistical/operational model of science, but it is not a necessary feature of scientific problem solving. A scientist will be working on a problem, and there may be a dominant theory and perhaps some other strong theories. The scientist designs experiments to test these theories. The results may be ambiguous, giving no clear falsifications of any of the theories, but returning even more questions. But this evolving set of questions is itself informative, increasing the 'problem content' (Popper 1974) of the phenomena that a successful theory must explain, thus constraining the theorising and imagination of the scientist. Eventually, if lucky, a scientist may propose a theory that answers a great many of the problems and questions that have been generated over the course of many research programmes and published papers, and the theory may then prosper at the expense of the old theories. It is not simply that scientists work within a 'paradigm' until anomalies become overwhelming and a new approach is taken, as in the philosophy of Thomas Kuhn (1962), but that scientists actively seek to solve problems, test theories, and explain data; the replacement of one theory by another may take place over a considerable amount of time. Within that time, experiments are generating more questions and clues, many of them flagged with easy-to-digest *P* values. This is in contrast to the piecemeal proposing and eliminating of hypotheses in operational research, where problem content does not grow and the discovery of new explanations is not the goal.

The image of falsification science as being no more than the iterative generation and elimination of hypotheses is an attractive one, in that it is closely analogous to the process of Darwinian natural selection, and therefore provides a powerful explanation of the evolution and development of good designs. This does seem to be the case for operational research, where models are proposed, tested, and changed, and tested anew. Criticism and falsification take on the role of 'negative feedback' in a cybernetic process in this vision, one endorsed by the Popperian and Nobel Prize winning scientist Sir Peter Medawar (Medawar 1969). The 'cybernetic' view of falsificationist *operational* research is, it seems to me, correct, but as a description of falsificationist *science* it is not so much wrong as incomplete, and potentially misleading. Criticism has another role in addition to the elimination of hypotheses, and that is the production of *questions*. By experimentation and other procedures for generating critical information, scientists occasionally manage to falsify a hypothesis outright, but nearly always manage to generate questions. As discussed above, the trial and error of scientific experimentation is as much about the accumulation of problems and questions as about the succession of hypotheses. The scientist then works in a way that resembles in many ways that of the detective, gathering clues that a good theory will have to explain. As questions accumulate they provide an increasingly detailed negative image of the explanatory content that a successful theory should provide. In some way, if only a little way, this explains how the trial and error process can inform the imaginative processes in which new theories are born. If a scientist poses a 'bold hypothesis' outside of this context, perhaps following the contradictory dictates of both classical science and operational research, then this is probably no more than what would be called a 'hopeful monster' in evolutionary biology, almost certain to be wrong in most of its particulars.

Big science, operational science, and the statistical model

Fisher was an agricultural researcher, but he was also an extremely bold theoretical scientist in the classical mode. Before Fisher, in 1909, William Sealy Gosset had already developed the t-test[5], again as a tool for operational research, in this case quality control checks in the Guinness brewery in Dublin. But it does not seem that the introduction of these methods gave rise to the operational research model of science. It seems more likely that it was due to the rise of 'big science' particularly during and following World War II. The understanding that organised research could lead to dramatic improvements in military and industrial capability led to a massive investment in science in the USA in particular. Big money and big research projects paid enormous technological dividends in the development of the atom bomb and the space programme – areas where operational research procedures rule supreme. Extremely complex experimental designs subject to computer analysis can be extremely efficient tools for industrial research, with its requirement to make engineering and manufacturing decisions. Procedures such as manifold adaptive experimental design have been extremely successful in engineering. Sequential experimental design in industrial production was so powerful that it was considered a military secret weapon in USA in WWII. But as tools in science, such experimental designs might sometimes hinder progress, as they are opaque to unaided human understanding such that critical analysis and understanding of the test procedure becomes very difficult, ruining the harvest of questions and problems that an experiment might yield. By contrast, great scientific theories may often be beyond the comprehension of the non-specialist mathematical scientist (especially in physics) but the experiments that test the theories may be of relatively simple conceptual design, potentially criticisable and understandable by anyone prepared to make the effort (this is not to say that the apparatus used in such experiments is simple.)

Where scientific understanding is required, big science has been in some estimation a failure. For instance, the 'war on cancer' initiated by Richard Nixon in 1971, despite billions of dollars invested, has yielded barely any dividend, with serendipity continuing to be the only source of new cancer therapies (Spector 2010). The non-scientist often asks, 'If they can put a man on the moon, why can't they find a cure for the common cold?' It is a good question, and I propose that the answer is that because the first is a question of operational research whereas the second requires the growth of scientific understanding. Big science answers big operational questions, but doesn't do so well with questions that require explanations. Big science, operational research, and explanationless science go hand-in-hand-in-hand. The values of classical science include truth, discovery, explanation, and understanding. The values of operational research are defined instead by the philosophies of pragmatism and instrumentalism, where explanatory theories are replaced by models – *instruments* that may or may not contain any truth or any explanation of things, but which simply *work*.

[5] Published under the pseudonym 'Student' (1908) and often known as Student's t-test.

Pluralism in science as a good

In the long haul, the two types of science must interact – technological progress depending on the growth of good explanations, and new theories depending on the growth of technology. The technological advances of today's *big* science are built upon the discoveries of yesterday's *great* science. If we are conscious of the differences in the two different approaches to science, explanationless science need not intrude beyond its domain, and indeed classical science can grow out of operational research. A researcher when trying to solve a problem of practical application may often encounter theoretical issues. To make progress, the applied researcher has to become a theorist, and in arriving at theoretical understanding, continue with a greater likelihood of solving the practical problem. These side-turns into theoretical science should be taken seriously and disseminated. It may be that work on practical problems is the fountainhead of theoretical science, with theoretical problems emerging from work on hard technical problems. Not only is operational research dependent on past scientific discovery, but scientific discovery may grow out of the problems uncovered by applied research. The moral of this is that not only should researchers beware an unconscious aping of the methods of other types of research when other approaches may be more fruitful, but also that individual researchers and research programmes could also benefit from a flexible, pluralistic approach to science. For example, operational researchers should be free to work more like 'curiosity-driven' scientists in the classical sense when the need or opportunity arises. By tackling challenging technical problems in the spirit of a quest for understanding as well as production, the practical people of applied science, industrial and agricultural research, engineering and design, can fertilise the traditional sciences with new problems and ideas, leading to theoretical advances, and true innovation. What is required is cross-fertilisation rather than competition between the various approaches to science – employing no single approach to the exclusion of others, nor giving in to the anarchic notion that in science anything goes.

References

Brooks, M. 2012. *Free Radicals: The Secret Anarchy of Science*. Profile Books.

Deutsch, D. 2011. *The Beginning of Infinity*. Penguin Books.

Feyerabend, P. 2010. *Against Method* (4th edn), Verso Books.

Fuller, S. 2003. *Kuhn vs Popper: The Struggle for the Soul of Science*, Icon Books Ltd.

Kuhn, T. 1962. *The Structure of Scientific Revolutions*. University of Chicago Press.

Le Fanu, J. 2002. *The Rise and Fall of Modern Medicine*. Basic Books.

Maggo, S.; Clark, D.; Ashton, J.C. 2012. The effect of statins on performance in the Morris water maze in guinea pig. *European Journal of Pharmacology 674*: 287–293.

Marks, H.M. 2003. Rigorous uncertainty: why RA Fisher is important. *International Journal of Epidemiology 32*: 932–937; discussion 945–948.

McGrayne, S.B. 2012. *The Theory That Would Not Die*. Yale University Press.

Medawar, P.B. 1969. *The Art of the Soluble*. Penguin Books Ltd.

Popper, K.R. 1959. *The Logic of Scientific Discovery*, Hutchinson Education.

Popper, K.R. 1963. Science: Problems, Aims, Responsibilities. *Federation Proceedings 22*: 961–972.

Popper, K.R. 1974. *Unended Quest*. Open Court Publishing Co.

Sinclair, J.D. 1988. Multiple t-tests are appropriate in science. *Trends in Pharmacological Science 9*: 12–13.

Spector, R. 2010. The War on Cancer: A Progress Report for Skeptics. *Skeptical Enquirer 34*.

'Student'. 1908. The probable error of a mean. *Biometrika 6*: 1–25.