Uncertainty and efficient science

1. Identifying science

For centuries scientists and philosophers have thought deeply about the nature of scientific research in an effort to make it more effective. The focus has been on justifying inferences and trying to find the most logically bullet-proof approach to building scientific knowledge. There has also been work to define what a scientific explanation is. Many interesting insights have been reached through this effort but also a lot of time has been wasted on dead ends. Meanwhile, various branches of science have progressed, some quickly and some not so quickly. For example, while physics, chemistry, and biology have raced ahead, psychology and economics seem to many to have lagged behind.

The emphasis on reliable knowledge is characteristic of science, but the approach taken in this article is to put the focus on (a) the wider goal of efficiency, and on (b) the exploitation of phenomena. This approach puts some very familiar facts about science in a fresh context and also suggests ways that scientific practice can be improved.

There is no single 'scientific method' and we cannot assume that everything done by people called scientists who work in laboratories is an example of science. Sadly, some of them have fabricated evidence, for example, which is fraud, not science.

So, how can we identify or define science? I suggest it be defined by the aspirations and methods of those who try to do it, whether they work in a laboratory or not. The details are what this article is about.

2. Focus on efficiency

The big picture is that science has been a highly effective approach to gathering knowledge and has helped to lengthen and improve our lives dramatically. As a society, we invest in it, first and foremost, because it enables us to improve our lives in practical ways, not just for the satisfaction of knowing things. As individuals, many of us use scientific methods with the intention of gaining valuable knowledge efficiently, not just to gain knowledge. Reliable methods reduce wasteful mistakes, so in that way promote efficiency.

This article focuses on methods for efficient science and does not present a major argument for efficiency. If you already think science should be useful then you do not need convincing. If you receive funding for useless research then there may be nothing I can say that will encourage you to think that is a bad thing.

But, just as some minimal support for the idea that science should strive to be efficient, consider these questions. Would physics be as well funded as it is without the electronics industry and nuclear power and weapons? Would physiology and biology be as well funded without medicine and agriculture? Would palaeontology and archaeology be funded at all without museums, television, and movies? Why should anyone be given a living for doing studies they find interesting but that will not help anyone else? Why should any group of academics be funded to publish papers to be read by others in their group but without ever providing any practical benefit to anyone who is providing the funding?

Useless research is done and funded, but much other research is useful or, at least, perceived to be.

Although it's too early to suggest an exact mathematical formula for the efficiency of science, it should be something driven by how much resource we, as a society, put into it and how much useful knowledge we get out. From this point of view, some obvious things that should be happening to promote efficient science are these.

- Focus science on topics that are likely to yield valuable knowledge easily.
- Do scientific research using efficient techniques.

These things already happen to a significant extent. Science is often done in laboratories using equipment carefully designed to generate and capture vast amounts of data accurately. Science funding is given to support research that is worthwhile. Private organizations that conduct research, such as pharmaceutical companies and market research companies, try to accomplish worthwhile work while controlling costs. Numerous further examples of methods that promote efficiency are listed later in this article.

Focusing on efficiency changes our approach to philosophizing about science because the most resource-consuming part of science is not the work of thinking about hypotheses (one of the major preoccupations of scientific epistemology over the past century or so), but is surely the work of making and recording many observations. For example, Charles Darwin almost certainly spent more time travelling, gathering specimens, and drawing them than he did on thinking up his famous theory of natural selection. That theory is so simple it could have been dreamt up in less time than it takes to draw accurately the beak of a single finch. The most expensive scientific studies are expensive not because of the work of inventing hypotheses and comparing them with data but because of the data gathering effort, including the equipment required. Particle accelerators and giant telescopes are good examples of very big investments in science and they are devices that allow observations to be made.

3. Focus on exploiting phenomena

How can this data gathering effort be directed at getting useful knowledge? It seems that in many cases, though not all, what scientists do initially is to characterize phenomena.

> *Example*: According to Gribbin (1984), when Max Planck first proposed a mathematical function that described the spectrum of black body radiation it had no physical basis[1]. He had simply found a mathematical way to combine two other equations, one that fitted the data quite well for low frequencies and another that fitted quite well for high frequencies. Physical interpretations of this new mathematical formula followed later.

These characterizations of phenomena can be used in practical ways and to build explanations of other phenomena, as discussed in more detail later in this article. What we think of as hypothesis development is usually thinking about how to characterize a phenomenon (e.g. choosing a mathematical

[1] The term 'scientific law' has often been used for mathematical characterisations of phenomena.

function that more accurately describes a pattern we see) or thinking of ways to explain some phenomena in terms of others (which might give us reasons for choosing particular mathematical functions that go beyond simple shape matching).

What is a phenomenon? Something as simple as a single crystal with an unusual appearance might be a phenomenon worthy of study for somebody, but more often phenomena are repeated observations of something that is, in some way, the same. If an experiment produces the same results each time then it could be a worthwhile phenomenon for study. If the experiment produces different results each time even when done in exactly the same way then the experiment has failed to produce a phenomenon worth studying. Simple, everyday physical systems are relatively easy to work with.

> *Example*: Hooke's Law is an equation that relates the length of something elastic (such as a spring) to the force acting on it. This is not an explanation of why a spring gets longer when you pull its ends, but it does describe something about the phenomenon of stretching, succinctly and accurately (within limits).

> *Example*: The persistence of physical objects is a phenomenon so fundamental that we can easily forget it. Counting is one way to characterize this and we rely heavily on counting and calculation. For example, if you counted the sheep in a field and then counted them again immediately and reached a lower total you would not assume that some of the sheep had simply ceased to exist. You would assume you had made a mistake in counting or that there was a hole in the wall.

> *Example*: Similarly, measuring allows us to characterize a wide range of phenomena that are predictable in some way. The dimensions of inanimate objects of many kinds will stay constant or vary under many circumstances and we know a lot about those.

While most phenomena of interest to science are repeatable or repeating phenomena, some one-off phenomena are interesting too. For example, what led to the extinction of dinosaurs? An increasingly detailed picture of the earth around that time and the changing populations of dinosaurs has been built up.

A characterization is a description, in words, pictures, mathematics, or some combination, of what has been observed. Characterizations of repeatable phenomena do not describe everything, but instead focus on what is essential, predictable, or perhaps unchanging. Characterizations of one-off phenomena tend to involve building an increasingly detailed and varied body of relevant facts.

What makes a good characterization? This too is a hard question but some are clearly better than others. Good characterizations of repeatable phenomena pick out something that remains the same, or is predictable. They are also succinct and accurate.

> *Example*: Boyle's Law is an equation that links the pressure of a gas to its volume, while temperature is unchanged. Again, it offers no explanation of why these variables are connected as they are, but it does describe a phenomenon related to gases. If refers to pressure and volume, which are predictably related and both essential to the phenomenon, but it does not refer to the shape of the container, or the time of day, or other factors that are irrelevant to the phenomenon.

Since the mid-20th century it has been typical practice in many branches of science to write papers in terms of 'hypothesis testing'. This somewhat obscures the extent to which studies have really been about characterizing phenomena, or where their lasting value has been through characterization.

In some cases the experimenter genuinely has one or more theories in mind and designs an experiment that produces results that agree with some theories but not with others. (This is often a way to show that someone else's theory is wrong and clear the way for your own.)

In other cases, the description in terms of hypothesis testing is probably little more than window dressing. The real intention is to extend the scope of an existing theory or characterization, or simply to see what will happen. As the years roll by and studies pile up on a particular phenomenon the finer points of the rival theories tend to fade into oblivion and review papers become catalogues of results.

Since hypotheses come and go it makes sense to design studies to also produce usable characterizations of what actually happens, regardless of what was predicted by theories under consideration at the time.

There is considerable scope for making science more efficient in future by better characterization of phenomena. Scientists often report their results with their attention focused on just the hypotheses of interest at the time. Consequently, they sometimes miss opportunities to create cleaner, fuller characterizations, fail to report details of what was held constant (but which others might want to know), and fail to make available their full data set to interested researchers.

Table 1 is an overview of the methods for efficient science discussed in this article.

4. Methods for efficiently characterizing phenomena

Here are some methods that have been efficient:

Laboratories: A laboratory is a place organized to facilitate efficient research. It typically has people, equipment, and access to data or objects of study (such as people willing to participate).

Apparatus and study paradigms: Most studies are conducted using an established set up such as an arrangement of apparatus or an experimental paradigm (i.e. a basic procedure for an experiment within which variations are possible). The set-up is designed to produce repeatable results conveniently and scientists like to use it more than once if possible. The results typically seen usually become the phenomenon of interest.

> *Example*: Once you have built a particle accelerator with equipment to measure the results of smashing particles together you naturally want to design lots of experiments that involve smashing different particles together and looking at what happens.

Table 1

Methods for efficiently characterizing phenomena

- Laboratories
- Apparatus and study paradigms
- Preference for predictability and simplicity
- Preference for surprising but consistent results
- Expansion
- Systematic variation
- Exploiting available learning opportunities
- Visualization with excellent information graphics
- Quantification
- Direct observation and nearly direct observation
- Holding on to individual differences

Using characterizations of phenomena

- Practical exploitation
    - Pick useful phenomena
    - See what you can do with what you have
- Use in explanations
    - See what you can do with what you have
    - Inferences from what is obvious
    - Simulation

Scientific evaluations

- Reliable procedures
- Statistical defence against coincidental results
- Tackling positive publication bias
- Continuous links
- Testing methods we use a lot but have not tested before
- Developing faster tests

> *Example*: Most psychology experiments are variations on established paradigms. The cycle usually begins when someone invents a new paradigm and publishes results. This gets replicated, then psychologists begin to argue over why the results turn out the way they do, conducting numerous variations on the original paradigm to try to tease out how

it works. In the process of these controversies a picture builds up of how the effect is driven by various different factors.

Preference for predictability and simplicity: Phenomena that are suitable for characterization tend to be predictable, regular, and easy to control. An experimental paradigm will try to control as much as possible and keep things simple so that the phenomenon can be seen clearly every time. If the study involves testing many members of a population and it turns out that the population is actually made up of two different homogeneous sub-groups then it makes sense to split the population and study each group separately so that a more predictable result can be achieved.

Preference for surprising but consistent results: Surprising results are more efficient because they provide more information. Being told something you already know provides no information, but a surprise is the opposite.

Example: The bystander effect is consistent and surprising. Many studies have shown that, when someone is in distress, the more people are around who could help the less likely it is that anybody will do so.

Expansion: Having established a repeatable phenomenon in one set of conditions, the pattern is usually to expand the range of conditions in which the phenomenon is characterized. Tweaks are made to the paradigm to explore the impact of new variables. Arguably this redefines the phenomenon itself.

Example: Robert Boyle's 1662 publication characterised the relationship between the pressure and volume of a gas with an equation but he kept temperature constant. Later, Charles's Law linked volume and temperature. Then, in 1834, Emile Clapeyron combined the two into the Ideal Gas Law, which relates pressure, volume, and temperature in one equation. Presumably the reference to an ideal gas reflected an understanding that real gases did not always behave perfectly in accordance with the Law. This would have been discovered by repeating the studies using different gases and more extreme conditions of pressure, temperature, and pressure. Though not captured in the Ideal Gas Law itself these departures from 'ideal' behaviour would have been clarified by the many similar studies performed.

A characterization that is fairly accurate within a set of conditions might be accurate in others too, or not. Sometimes the boundaries beyond which a characterization works are known, and sometimes they are not.

Researchers typically seek to expand their characterization of a phenomenon by (a) considering additional variables, usually independent variables, and by (b) considering values for independent variables that are outside the range previously studied.

In choosing variations on a paradigm to expand the scope of the phenomenon that has been characterized the aim is usually to expand it successfully, and perhaps find the boundaries beyond which the characterization is not accurate. Scientists usually do not look for variations so different from their previous experience, and so much more complex than previous studies, that they have little chance of relating what they observe to what they have observed in the past.

When expanding the conditions in which a characterization applies, the focus is often on finding the difference between what the old characterization predicts and what the new results show.

Systematic variation: Paradigms usually involve a number of independent variables the experimenter can manipulate, and a number of dependent variables, which measure things that are supposedly driven by the independent variables. If possible, it is usually efficient to vary the independent variables in many small increments individually and in combinations, and use graphs and mathematics to show how independent and dependent variables move together. If an effect holds incrementally it is much more convincing than just using two settings for each independent variable.

Example: Roger Shepard and Jacqueline Metzler (1971) showed people pairs of pictures of shapes rotated to 10 different extents and found that the time to decide if the shapes were the same or different was a linear function of the angular difference between the pictures. It was as if people mentally rotated one picture to decide if they were a match. If they had just compared two levels (e.g. no rotation versus 90 degrees) the results would have been far less arresting.

Example: Fitts's Law relates the time it takes to move your hand to hit a target to the distance covered and the size of the target. In the original paper (Fitts, 1954) Fitts reported three experiments where the tasks were similar but different (and were variations on studies by others before him). The first experiment gave results for four different distances, four different target sizes, and two different stylus weights, making 4 x 4 x 2 = 32 different conditions. The second experiment was 4 x 4 = 16 conditions. The third experiment gave results for 4 x 5 = 20 conditions. The resulting data would have made impressively clean graphs, but the paper lacks them, relying on statistics alone. (When I plotted the figures for one of the studies I noticed a possible reason why graphs weren't shown originally, which is that systematic departures from Fitts's formulae can easily be seen!)

Exploiting available learning opportunities: The gold standard for research is the controlled experiment, where we manipulate some variable and see what effect that has. However, sometimes this is hard or impossible to do, so we also use other methods, often exploiting learning opportunities that arise naturally. These occasionally amount to experiments, but usually involve using data already collected and searching for statistical regularities that might be signs of causal links. This strategy is not always successful.

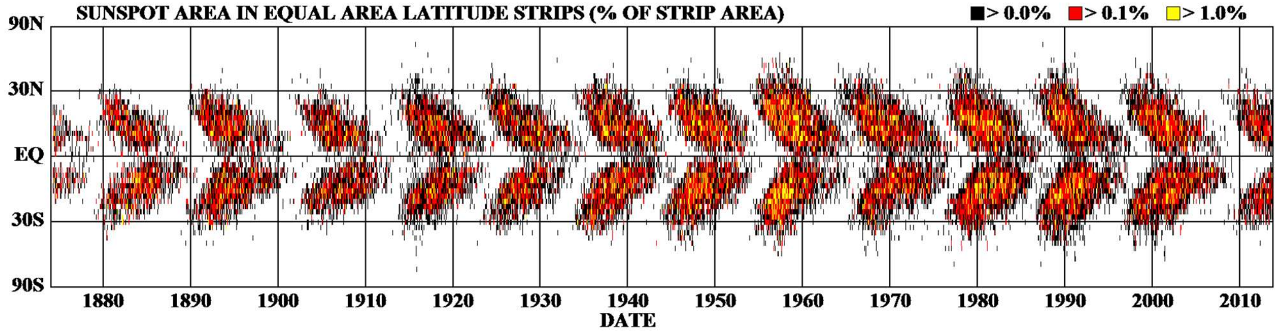
Example: There have been many regression studies of corporate governance practices using large samples of companies to see if the corporate governance practices they report link to accounting restatements or profits. These have been inconclusive, probably for a number of reasons. It would be more effective now to concede that crunching multi-company data sets is not working and set up closely monitored interventions within a small number of organizations to identify any observable immediate and knock-on effects of corporate governance changes from the moment they are first discussed to the 5 year anniversary of their introduction.

Visualization with excellent information graphics: If relationships between variables can be shown in nice graphs this can help greatly in finding order within data. Systematic variation helps with this.

Example: Edward Maunder's Butterfly Diagram (Figure 1) shows how the location of sunspots varies over time in both the northern and southern hemispheres of the sun. It was published in 1904.

Quantification: Rather than just say that there is 'an effect' or that a variable is higher in one condition than in another, it helps to quantify a phenomenon by fitting mathematical formulae. If necessary, these can include probability distributions to characterise variations.

Figure 1: Modern version of Maunder's Butterfly Diagram by NASA.



Quantifying phenomena is standard practice in physics, but surprisingly rare in psychology. Countless studies have explored factors that make learning easier but if you want to know how long it will probably take to learn something it is unlikely that you will be able to find a study that tells you that. If physicists did psychology then that would have been sorted out decades ago.

Example: Quantum physics is a body of science concerned with the behaviour of light and very small things like atoms and electrons. It is largely expressed using mathematical equations that fit the results of experiments and do a good job of predicting the results of more. The area is so mathematical that physicists talk about alternative "interpretations" of the equations, which are what most people would see as explanations. For example, Heisenberg and Schrödinger produced different equations that turned out to be mathematically equivalent. Heisenberg had started out trying to focus on measurable quantities but thought of electrons as particles. Schrödinger thought of electrons as waves and used equations established to describe waves on water (Gribbin, 1984).

Direct observation and nearly direct observation: One thing that makes science much, much easier is being able to see things happening as if in real time. Inventions like the microscope, the telescope, high speed film, time lapse photography, and the functional MRI machine have allowed rapid progress.

Example: Neuroscientists spent decades speculating about what bits of the human brain might do, relying mostly on studies of people with brain damage. Once the fMRI machine was available they could watch the inside of the brain in real time and in 3D as healthy people did tasks given to them by the neuroscientists. This was a breakthrough.

Example: Sometimes, just getting close to direct observation can be very helpful. In the early 20th century nobody had seen a molecule but long before that a British botanist, Thomas Brown, had reported seeing that a tiny grain of pollen floating on a droplet of water appeared to jiggle about when viewed under a microscope. Einstein suggested this was because it was being buffeted by many moving particles and showed that its random movement agreed with that interpretation (Gribbin, 1984).

Holding on to individual differences: Studies of populations of individuals are more valuable if details of each individual are used rather than just being pooled into average results.

Example: Many studies of human reasoning have shown that, on average, we make a number of predictable reasoning mistakes. However, when differences between people were studied more closely it emerged that some people are less prone to them than others (Stanovich and West, 1998). For example, which food is healthier, the one described as '97% fat free' or the one described as '3% fat'? If you are one of those people who mentally translates '97% fat free' into '3% fat' to guard against this kind of marketing then you have experienced one way that such individual differences can arise.

Example: Similarly, it has emerged over the past two decades that some people respond to over-eating and physical exercise very differently to others. When overfed, some people get fat, though to varying degrees, while a few others put on muscle instead. In Bouchard et al (1990) 12 pairs of identical twins were over-fed and under-exercised to the same extent for 100 days. Visceral fat gains (the most worrying kind) ranged from 4 cubic centimetres to 46 cubic centimetres. Those gains also correlated strongly between the twins. Similarly, some people benefit from physical exercise much more than others on particular measures of health and fitness. Some people can improve their VO2max by 50% through exercise, while others who make the same effort achieve no increase at all (HERITAGE Family Study research consortium, 2012). This is a complex area and you could be a non-responder on one measure but still benefit from exercise overall.

It can also happen that individual curves of one shape, when averaged, become a curve of another shape.

5. Using characterizations of phenomena

Once phenomena have been characterized, even if they cannot be explained, it is possible to make use of them in at least two ways:

Practical exploitation: If we know that some manipulations give particular results, we can start to exploit that in practical ways, such as in the design of machines, teaching methods, and health programmes.

> *Example*: If you know that the efficiency of an engine depends on the temperature difference in it then you can start to think of ways to raise the maximum temperature and improve cooling elsewhere.

> *Example*: The equipment used by early scientists investigating electricity included components that then developed into the components used in practical circuits, such as chemical batteries, coils of wire, and variable resistors. It is very hard to draw a dividing line between science and technology in this example.

Use in explanations: Many explanations of phenomena are in terms of other phenomena, ideally ones that have already been well characterized. It is possible that all satisfactory scientific explanations are phenomena explained in terms of other phenomena.

> *Example*: Why does hot air rise? One explanation is in terms of gases expanding when warmed, thus becoming less dense, and then the less dense body of air floating above the denser air around it. This explains the rising air in terms of three other phenomena: expansion with temperature, reducing density with expansion, and floating. This feels like a proper explanation.

> *Example*: Another explanation that goes beyond just characterizing a phenomenon is the Kinetic Theory of gases. In this explanation, the behaviour of gases is explained in terms of the statistical properties of many molecules, each moving at high speed, colliding with each other and with objects. High temperature is explained as the molecules moving faster.

> *Example*: Explanations of one-off phenomena can also be created. The extinction of the dinosaurs might be explained in terms of asteroid impact, atmospheric changes, and so on.

Provided you stay within the conditions within which a characterization is thought to be adequately accurate, predictions are usually correct and can be useful. (This does not mean there is no value in further research into other conditions, or into explanations.) Similarly, if an explanation is known from previous studies to work well within certain conditions then that explanation and the predictions it makes can be useful for predictions. Extrapolated beyond previous experience, the explanation could fail and be misleading, but it is likely to be better than guessing.

The tendency for characterizations and explanations to be more reliable in situations very like those used in past studies, but less reliable otherwise, gives a reason for focusing on phenomena with obvious practical applications.

6. Methods for efficient practical exploitation

Some methods for improving the efficiency of practical exploitation of characterizations of phenomena are these:

Pick useful phenomena: Sometimes a distinction is made between ‘pure’ science and ‘applied’ science, but there are times when a phenomenon with obvious uses can be studied easily and yields results that generalize well. What is studied need not be ‘natural’. For example, it could be the behaviour of electronic components, or the swirls of air within a vacuum cleaner, or features of multimedia teaching materials.

> *Example*: The Stroop Effect was named after John Ridley Stroop, who published the first English-language paper about it.

People are asked to look at words printed in different colours and say the names of the colours as quickly as possible. If the word spells one colour but is printed in another colour then people tend to take longer to do the task and that extra time is the basic Stroop Effect. Stroop's paper is one of the most cited in the history of psychology with over 700 repetitions of the study published (MacLeod, 1991) and countless more done by students.

The pattern of replications and variations is a great demonstration of how characterizations of phenomena are expanded, but a poor demonstration of exploitation. Although it could have been done more efficiently, the real reason why this has been such a wasteful line of research is that the Stroop Effect has so little practical use. Is there really much danger of gadget or software designers presenting colour information using conflicting words? It could happen, in which case science can offer over 700 published studies showing it to be a bad idea, but that's not much of a practical payoff.

Example: In contrast to the useless Stroop Effect, the Brown-Peterson release of proactive inhibition effect has obvious applications in education. This effect arises in the Brown-Peterson task, which involves remembering digits or letters for a few seconds while doing a distracting task. Memory on the first one or two trials is very good but soon drops if the task stays with digits (or with letters). If what has to be remembered is changed from digits to letters, or vice versa, memory improves again for a couple of trials. Does this have anything to tell us about how to teach mental arithmetic to children? It should do. If they get confused doing mental maths then give them a short break and try again.

Example: During the mid-20th century psychologists interested in human memory focused a lot of attention on one task: learning a list of random words. Because this is almost never something people need to do in real life the results of those studies are of little direct use in education. How much more useful it would have been if the same effort had gone into learning to spell new words or memorise true number facts. These are things we spend many years teaching children, sometimes with dismal results.

See what you can do with what you have: Rather than hunting for a phenomenon that helps you do something you want to do, it is sometimes more efficient to start with a phenomenon you can control and look to see what you can do with it.

7. Methods for efficient use in explanations

Once phenomena have been characterized it may be possible to use them to explain other phenomena. For example, having seen how one shopper is influenced by the purchases of his or her friends, we might try to explain phenomena we have observed in the behaviour of large groups of shoppers, or perhaps predict some group behaviour and see if it happens.

Again, there are relatively efficient ways to do this.

See what you can do with what you have: Rather than hunting for the explanation of a target phenomenon, it is sometimes more efficient to look for phenomena that might be explained with a phenomenon that has been characterized.

Inferences from what is obvious: A strategy that can be very powerful is to deduce and predict as much as possible from what is obvious and then work to explain any differences between those predictions and reality.

Example: Darwin's idea of 'natural' selection is one that is so compelling that it is obvious that it must happen to some extent. Artificial selection was a well-established practice and it is obvious that some individuals produce more offspring than others, often because they survive long enough. Natural selection simply has to be operating to some extent. What remained to be established was just how much of the evolution of species could be explained by this phenomenon alone.

Simulation: Predicting phenomena from other phenomena can sometimes be done by mathematics, but today it is often easier to do it by computer simulation. This also helps to avoid the mistake of focusing only on models that are mathematically easy to work with.

8. Scientific evaluations

Scientific evaluations of such things as new drugs, social programmes, and management methods lie on the border between science and technology. Reliable results are usually the main focus and many familiar techniques are used for this. However, efficiency is also important. Here are some good techniques to use.

Reliable procedures: These include control groups, double blind designs (where neither the subjects involved nor the experimenters know who is getting each treatment), random assignment to groups, and avoidance of leading questions in interviews and questionnaires.

> *Example*: A number of attempts have been made over the years to show that implementing Enterprise Risk Management (ERM) is good for businesses. These illustrate failure to evaluate scientifically because they rely on correlations rather than experiment. One type of study shows that using ERM is linked to high profits, or market value, but this could be because companies easily making lots of money, or with very capable management teams, have time to waste on regulator-approved schemes like ERM. Another type of study shows that companies who have recently appointed a Chief Risk Officer subsequently do a bit better. This might be because companies appoint CROs most often when they can afford to, or perhaps just after something has gone wrong and they wish to reassure investors and regulators. By reversion to the mean, companies that have just had a very bad year will usually have an easier time the next year, which makes the CRO look good even though it probably would have happened anyway.

Statistical defence against coincidental results: Significance tests, confidence intervals, and Bayesian procedures exist to help distinguish between results that are just unlucky sampling and results that are real.

Tackling positive publication bias: There has been a tendency only to publish the results of tests showing something has worked, or where some interesting result was observed. This left important evidence unpublished. Procedures are now used to try to reduce this problem, mainly by requiring studies to be registered before they are performed and then published no matter what the result. There are also restrictions to prevent changes to the procedure or analysis methods after initial design. Another route has been to provide journals specifically for boring papers with no interesting differences found.

Continuous links: Finding a continuous link between, for example, the extent of a physiotherapy and its beneficial effects, is more convincing than just finding that those who had the therapy at one level of intensity recovered better than those who did not have it at all. A continuous link like this is also a better characterization of the phenomenon of therapeutic effect from the treatment.

Testing methods we use a lot but have not tested before: Tests of previously untested methods can have a stunning effect and may be inspired by realising that a widely used method is illogical and should perform poorly, or by realising that people seem to be ignoring obvious signs of poor performance.

Developing faster tests: Scientific testing can be made more efficient by developing tests that can be applied more quickly.

> *Example*: Testing if changes to diet affect cancer risk is extremely slow and uncertain work. However, if it is established confidently that DNA methylation is part of the mechanism of cancer then tests of DNA methylation can be used instead of cancer incidence to gain feedback on which drugs, dietary changes, and exercise regimes actually work.

9. Hard and easy topics for science

With ideas about how to do efficient science in mind it is easier to see why some scientific projects are harder than others. Here are some well-known problem areas.

Health risks: The battle to work out the elements of a healthy lifestyle is made harder by the complexity of the human body and the fact that many of the health effects of interest happen rarely, and unfold over many years. The most controlled situations are on tissue samples in laboratories and using animals, but even these are complex systems compared to a piece of metal or a bottle of gas. In addition, even when a reliable phenomenon is isolated in one of these laboratory paradigms it usually proves very hard to find it in people and show that it has an impact on health long term.

Studies to do this tend to rely on correlation rather than manipulating behaviour to see what results are produced. The correlations are hard to interpret because there are so many measured and unmeasured factors that could be involved and they often correlate with each other. With a big enough sample, lots of statistical 'links' will be found, but what they mean is hard to say.

Hopefully, the development of tests such as for DNA methylation will accelerate research greatly.

Quantum physics: Work to explain physical phenomena in terms of entities no bigger than atoms has produced some very accurate mathematical characterizations of phenomena and a long argument about whether they are something more than that. Are the entities named in models of atoms real objects or just notions that interpret the equations? When atoms are smashed, are the particles that fly out from the collision bits of broken atom, or something else?

I suggest that quantum physics is hard because of a combination of two main factors.

First, the aim is to build explanations of phenomena we have characterized (at the observable, measurable level) in terms of phenomena we have not characterized but can only imagine, or, at best, guess from the behaviour of fragments of atoms that have just been smashed. This is much more difficult than the usual scientific task of building from well-known, independently studied phenomena.

Second, some of the phenomena to be explained are so bizarre it is difficult to think of any known phenomenon that might account for them. Spooky Action At A Distance is most famously demonstrated by the double slit experiment with very slow release of particles towards the slit, and by Bell inequality experiments using 'quantum entanglement.' How can anything produce this effect? There simply is nothing in our ordinary experience of the world that can account for it, so physicists invent phenomena (e.g. particles, forces, properties) that suggest equations that seem to work.

I get a similar feeling of bafflement when wondering how a gyroscope is able to tilt over without falling. Nothing I know about physics, either intuitively or from school learning, can begin to explain it.

Some have argued that, since the equations do a good job of describing behaviour then the entities they sort of refer to must exist. This is not very convincing, however, once you have seen symbolic regression in action. This is a technique where a computer tries fitting lots of equation forms to data, gradually finding better fitting structures (not just tuning parameters values). It seems obvious that thousands of physicists trying thousands of equations might produce equations that fit really well by massive trial and error, having no real idea why the equations work so well.

Cognition: Scientific study of cognition has been a struggle. The great Herbert A Simon proposed one promising strategy of trying to understand a small set of basic information processing operations and then building models to simulate performance of more complex tasks using those atomic units. Although some good work was done and some very impressive simulations have been created, this approach did not create the huge progress that was hoped for. Characterizing those atomic information processing operations proved tricky because they could not easily be isolated. Also, even when a simulation did a good job of modelling actual behaviour, how could it be used?

10. Inefficient practices to be avoided

Some branches of science could become more efficient by eliminating common but wasteful practices such as those listed in Table 2 and discussed below.

Focusing on statistical significance at the expense of quantified characterization: This is typical in psychology, where countless papers have been published with 'significant differences' and review papers assess the strength of effects in terms of how statistically significant they are, without considering their actual size. For these meta-studies a statistic called Cohen's d is often used, which is the difference in means divided by the standard deviation.

> *Example*: Capeda et al (2006) provided what they call a quantitative review of the results of 839 experiments on distributed practice for verbal learning without producing any equations to show how much time can be saved by distributed practice.

Many opportunities for good characterization of phenomena have been missed this way.

Table 2

Inefficient practices to be avoided

- Focusing on statistical significance at the expense of quantified characterization
- Confusing statistical and theoretical hypotheses
- Fixating on the search for an explanation
- Expensive studies that only show links
- Studying artificial systems as if they are natural systems
- Dredging heterogeneous cases with regression methods
- Statistical personality constructs
- Incomplete reporting

Confusing statistical and theoretical hypotheses: Many business research studies today confuse statistical hypothesis testing with theoretical hypothesis testing. The researcher lists 'hypotheses' that are nothing more than predictions about whether one number will be bigger than another, often with no clear theory for why that prediction is made.

Fixating on the search for an explanation: A familiar pattern in psychology is one where an initial paper presents an intriguing phenomenon and then inconclusive controversy rages for years as to why that phenomenon occurs, with little or no attempt to exploit the phenomenon or use it to explain others.

Expensive studies that only show links: Some studies designed to search for correlations between variables are time consuming and expensive. Very often the results are equivocal and this was an obvious problem from the start.

> *Example*: Can eating chocolate help to keep you slim? In a typical correlation study (Golomb et al, 2012), 1,018 people (70% male) filled in a questionnaire about their eating and exercise habits and were weighed. Those who ate chocolate five times a week had a lower BMI, on average, than those who did not. The quantity of chocolate consumed was not related to weight. Although the idea of slimming by eating chocolate is immensely attractive to many people, the results could equally well have been stated as showing that fat people avoid chocolate, or at least say they do. The article has lots to say about reasons why chocolate might help you stay slim, though the researchers concede that they only have a correlation and need to do an experiment (i.e. change the chocolate intake of people to see what happens) to go further. This was obvious from the start. If there is a useful effect of chocolate then it will be large enough to show up with far fewer than 1,000 people. Would it have been so much harder to recruit, say, 100 people to eat specified amounts of chocolate and be weighed repeatedly than it was to recruit 1,018 people to be weighed just once?

Studying artificial systems as if they are natural systems: By 'artificial systems' I mean systems created by people. It is obvious that, for example, electronic components, the body shapes of racing cars, and the layout of road systems are all artificial.

The way to study them is to combine tests of alternative designs with tests of elements within those designs that you hope will reveal controllable phenomena that can be exploited in future designs.

Artificial systems more often mistaken for natural systems include economies, societies, and most of our cognition. To illustrate how clear cut this can be, consider the way people solve quadratic equations. Obviously, it has a lot to do with the way they were taught to solve quadratic equations and is not really a 'natural' phenomenon, like water freezing or apples falling from trees. Similarly, the way the UK government responds to economic changes is at least partly the result of its theories about how economies work and the techniques for control it has developed. Finally, the way peaceful revolutions have occurred in recent decades is partly the result of peaceful revolutionaries who have developed and written guides to their techniques for peaceful revolution.

Studying things that people invent, whether it is the body shell of a racing car or tactics for nonviolent revolution, requires appropriate methods.

An appropriate way to study cognition is to test alternative ways of thinking through the same task, and alternative ways to develop the knowledge needed for those alternative methods. We can also study elements of those methods to see how their results can be varied, hoping to characterise controllable phenomena that can be exploited in future designs.

Another useful approach is to study the thinking of a set of different individuals facing the same task and look for links between the tactics each uses and the results they obtain. The final step is to teach the best tactics – the ones used by the high performers – to the low performers to see if their results improve.

With economies and societies it is not so easy to implement alternative designs for the system as a whole and even testing alternative ways for a person or organization to behave within an existing society/economy can be difficult because participants may be risking their livelihoods. However, research techniques that may be appropriate include these:

- Testing designs on a smaller scale, in a simplified situation (e.g. a group of people in a laboratory carrying out a negotiation task).
- Studying what actually happens when people/organizations behave within an existing society/economy. This might reveal if the consequences of the design are what was intended, or somehow different.
- Studying the impact of changes to rules that control an economy/society by looking to see what actually happens and trying to link this to the changes made.
- Looking at differences between societies and economies that exist now (different countries, tribes, other social groups) to see if the way they are designed (the rules, structures, processes, etc) can be linked reliably to the results they obtain (longevity, education, happiness, etc).
- Studying the way people create societies/economies to see if some approaches result in societies/economies that provide more peaceful, happy, prosperous lives than others.
- Using computer simulations of artificially intelligent actors interacting within simulated societies/economies.

With the active participation of governments and their agents it might be possible to do even better than this by organizing properly controlled experiments on a massive scale. (This is instead of what usually happens, where governments guess which rules will lead to improvement and introduce them for everyone at the same time, preventing controlled comparison.)

A lot of research resources have been wasted in psychology because of not trying to control what participants think while doing experimental tasks. This may have been due to the mistaken belief that minds are natural, like a test tube of chemicals or a piece of rock, but there have been other justifications too. People are rarely asked to think in particular ways because this is an unreliable approach; some psychologists have gone so far as to say that only observable behaviour is important.

In reality we know that what people say about their thinking is frequently unreliable, but sometimes agrees with their behaviour and performance. Also, what people are asked to do is not necessarily what they actually do, but instructions can be effective and can produce behaviour that is different even though the task is not.

More recently, machines able to see which parts of our brains are active while doing mental tasks have shown that giving people different instructions about what to think, or how, can produce different patterns of activation. Deliberately trying to think in particular ways is a meaningful and useful thing to do in science.

What this means is that what people say about their thinking can be useful but needs to be used with caution and, as far as possible, their claims need to be checked against other data.

Dredging heterogeneous cases with regression methods: Another over-used method has been to grab big data sets and hit them with regression equations featuring large numbers of explanatory variables. Very often the results are unsatisfactory and unconvincing, with different results from slightly different samples. I think one of the reasons for this is almost certain to be that most of the populations involved contain very different cases and the mechanisms in each sub-population are different. The regression picks up very slight statistical tendencies but they are not necessarily meaningful in any of the sub-populations.

A better strategy would be to combine the search for links with a search for clusters. Purely statistical analysis may not be the best way to do this; human knowledge may be needed to guide the algorithms.

Statistical personality constructs: A research strategy that has been used hundreds, probably thousands of times, goes like this. First, think of a long list of questions that you imagine will be answered differently by people who have different personalities, or attitudes, or beliefs. Next, get hundreds of people to answer those questions. Then apply a statistical method called factor analysis to find groups of questions whose answers seem to correlate quite highly among the people who did your questionnaire. Give those groups of questions names that conjure up the sort of person you think would tend to give those answers. Finally, to demonstrate the 'validity' of your research, get people to do the questionnaire again a few weeks later to check that they give similar answers each time, and see if the numbers from your statistical factors correlate with behaviours you think are relevant. If they do, claim this supports the validity of your constructs.

In principle this is another example of trying to explain one phenomenon (e.g. the tendency to participate in dangerous sports) in terms of another (i.e. the scores on factors that you get by having someone complete the questionnaire). So, why does this seem so unsatisfactory to me?

Part of the problem is that the statistical constructs (i.e. the factors) do not represent particular beliefs, or strategies, or physiological characteristics, or skills. Sometimes the statistical pattern exists because of those, but the construct itself is free of such familiar material. To me it seems obvious that we would be better off focusing on beliefs, strategies, skills, or physiological characteristics directly. These are richer, better understood, and allow predictions that go far beyond weak statistical predictions.

I also worry that the factors you find may simply reflect the questions you asked in the first place. It is true that questions that don't seem to correlate with any big factor are dropped from these questionnaires, but that does not provide reassurance that all relevant questions have been asked. Maybe those questions left out could be put back in if we just invented 10 more questions that ask almost the same thing and try again.

Another part of the problem is that the models created by this statistical strategy typically do a poor job of predicting behaviour. Usually, if you want to know what someone will do in future, you will get better results by just asking them their intentions than you can by asking 100 questions about their personality.

The strategy of fishing for statistical personality constructs makes it to this list of methods to avoid because there are alternative methods available that are more familiar and more effective.

Incomplete reporting: Some reports of scientific studies fail to give all the information needed for their results to be understood and used in quantitative characterization of the phenomena involved. For example, they may give incomplete explanations of quantities measured, or how those were summarised or transformed. They may give too few details of the rest of the paradigm, focusing instead on the independent and dependent variables only. They may collect just a few data when they could, almost as easily, have collected far more.

11. Scientific evaluations of scientific methods

Scientific evaluation of scientific methods gives us a way to decide what is good science. We are looking for methods that produce useful, reliable knowledge, but quickly and cheaply. A variety of formulae might be used to capture this idea, along with a variety of experimental paradigms.

This approach would confirm that rationality is not purely an arbitrary and cultural matter, because some methods and some inferences work better than others.

Unfortunately, scientific evaluation of scientific methods is surprisingly rare. Evaluation is very often thought about, of course, but competitive tests of alternative methods to see which perform best are rare.

Some relevant research includes the following:

Testing statistical methods: It is very common to test statistical methods on real or synthetic data to see which methods work best.

Exploring biases and failings of the scientific process: Studies have also been done to look at biases such as experimenter effects, statistical biases, positive publication bias, and placebo effects, and to study failings of peer review, and the tendency to ignore rebuttals, and even retractions of, flawed papers. Some studies have also been done to test ways to eliminate these problems.

However, these generally do not compare alternative research methods to find out which work best.

Simulated science in the laboratory: A more general type of test uses a simulated scientific research task. Mynatt et al (1978) created a simulated universe and challenged people to discover its laws. Their study showed that students were generally abysmal at uncovering the laws of the imaginary universe, and actually performed slightly worse when given training in a form of hypothesis testing.

Similarly, Dunbar (1993) challenged subjects with a simulated research task involving genes, and Schunn and Anderson (1999) asked people to study why distributed learning is more efficient than massed learning.

In these studies the objective was to understand 'scientific thinking' as if it is a natural phenomenon. There was no direct attempt to test alternative strategies. However, these simulated tasks show how alternative strategies might be tested.

Objective, scientific tests of the performance of research methods might help to settle the long-running battles over qualitative research, especially methods from sociology often known as interpretivism. My feeling at present is that objective testing of methods like Grounded Theory would show them to be next to useless as well as time consuming.

To illustrate the issues, consider psychoanalysis in the first half of the 20th Century. An interpretivist who talked to patients and analysts at that time would have heard from both that the process was useful and that some patients at least got better as a result of it. The interpretivist would have gathered numerous statements about this, why the method is effective, why people do it, what it means to them, and so on. All this would have been summarised systematically and presented as an (internal) theory of psychoanalysis, 'grounded' in the statements of the participants.

The trouble is that nobody actually knew if the treatment was helping. Eysenck (1952) summarised published literature on recovery from neuroses and reported that 64% of patients who received 'eclectic' psychotherapy (i.e. a mix of things but not including psychoanalysis) recovered while only 44% of patients receiving psychoanalysis recovered. The 64% recovery under 'eclectic' is about the same as for patients receiving no therapy. He wrote that it was difficult to establish definitely if the patients in the various groups were more or less seriously ill when they started but close reading of the various research reports suggested that there were no significant differences.

As a treatment, psychoanalysis turned out to be useless, or perhaps worse than useless, and the only people to benefit consistently were the psychoanalysts who gained a livelihood from their procedures.

A more modern issue along similar lines is the controversy over the value of immediate psychological counselling for people who have been in a traumatic event. Most people take it as fact that 'talking about it' helps but controlled research does not consistently back this up and some results suggest that talking about it makes things worse (e.g. Rose et al 2003). Other types of support and therapy are probably more effective and efficient.

Another example of a situation where interpretive research methods could be misleading is in tackling the problem of teaching arithmetic to children. Today we spend a huge amount of time teaching arithmetic to children and yet by the age of 11 many still struggle. There are 328 basic number facts to learn so this suggests that, for some children, spending over 3 hours on each fact distributed over several years, is not enough – at least not with today's teaching methods.

Suppose research was conducted into the reasons for this slow progress for many children. The interpretivist would spend much time talking to students, teachers, teaching assistants, and parents. From this a summary of the points made by all those people would be created.

I think it is inevitable that, from the children and their parents, some frequent themes would be that:

- Maths is confusing.
- Maths is boring.
- Teachers do not make maths interesting.
- Maths is useless anyway because we can use a calculator.

For children suffering the emotional pain of persistent failure to learn, these ideas are attractive for obvious reasons. They help lessen the pain.

Teachers and teaching assistants are likely to mention themes that include the following:

- We don't have time because of all the bureaucracy.
- The government keeps changing its instructions.
- Parents don't do enough at home.
- We're actually doing a very good job because most students have mastered arithmetic by the time they leave primary school.

Again, there are obvious reasons why teachers would find these ideas plausible and want to mention them to anyone researching the problems of arithmetic teaching.

Nobody knows why some children struggle so much with arithmetic and we need reliable research to reveal this. However, compiling the statements of people involved is not, on its own, likely to establish anything useful beyond a very clear picture of how people currently choose to defend themselves when the topic is raised.

The same sort of issues would arise if investigating the performance of risk management and corporate governance methods. Terrible mistakes still occur, but most people involved have good reasons for talking up the effectiveness of the methods they have promoted, or the effectiveness of what they have done. Just summarising what people say is unlikely to provide new, helpful insights or discoveries.

As a final example of a topic that would most likely defeat interpretivists, consider the topic of 'risk attitude'. The idea that we all, innately, have an attitude to risk, and that this strongly drives our choices about behaviour, is widespread. If you just ask people about their 'attitude to risk' they will give you answers that reflect that idea. Almost nobody will say that they don't have one. Risk taking and risk attitude are taken as virtually the same thing, though of course an attitude to risk (if it exists at all) is not the same as risk taking.

In fact, the extent to which we perform risky acts depends on how risky and rewarding we think they are, among other things (Schoemaker, 1993). It is not possible to infer differences in risk attitude simply from differences in risk taking, even for objectively identical situations. Asking people how risky they think a behaviour is tends to explain at least some of risk taking differences (e.g. Weber et al 2002), and there are many more factors that can be included. Research shows that, if people have a risk attitude at all, it is not consistent between different situations.

Because reality is so different to popular beliefs, an interpretivist is likely to come away with just a very nice summary of popular beliefs. In particular, asking people about their 'risk attitude' will prompt them to talk as if they have one, even though this is by no means certain. Risk attitudes may not exist at all (Chater et al, 2011).

In summary, testing research strategies scientifically could eventually reveal which are most efficient. That is, which are most likely to find the knowledge that leads to methods that protect people from long term psychological trauma, make arithmetic a joy for more children, and provide the means to control risk taking in organizations.

12. Summary

The suggestions above amount to a simple recipe for efficient scientific progress. It involves searching for and describing regular relationships between things we can do and results we want, expanding the range of what is known and predictable, gathering data efficiently, and exploiting that knowledge. It does not involve interminable controversies over the causes of behaviour that is of no practical use.

13. References

Bouchard, C., Tremblay, A., Després, J-P., Nadeau, M. D., Lupien, P. J., Thériault, G., Dussault, J., Moorjani, S., Pinault, S., and Fournier, G., 1990. The Response to Long-Term Overfeeding in Identical Twins. *New England Journal of Medicine*, 322 p.1477-1482. Available at:<http://www.nejm.org/doi/full/10.1056/NEJM199005243222101#t=articleTop>.

Capeda, N. J., Rohrer, D., Pashler, H., Vul, E., and Wixted, J. T., 2006. Distributed Practice in Verbal Recall Tasks: A Review and Quantitative Synthesis. *Psychological Bulletin*, 132(3), p.354–380. Available at:<http://uweb.rc.usf.edu/~drohrer/pdfs/Cepeda_et_al_2006PsychBull.pdf>.

Chater, N., Johansson, P., & Hall, L. (2011). The Non-Existence of Risk Attitude. *Frontiers in Psychology*, *2*, 303.

Dunbar, K. (1993). Concept discovery in a scientific domain. *Cognitive Science*, *17*(3), 397-434. Available at: http://onlinelibrary.wiley.com/doi/10.1207/s15516709cog1703_3/pdf

Eysenck, H.J. (1952). The effects of psychotherapy: An evaluation. *Journal of Consulting Psychology*, 16, 319-324. Available at: http://psychclassics.yorku.ca/Eysenck/psychotherapy.htm

Fitts, P. M., 1954. The information capacity of the human motor system in controlling the amplitude of movement. *Journal of Experimental Psychology*, 47(6), p.381–391. (Reprinted in Journal of Experimental Psychology: General, 121(3):262–269, 1992). Available at: <http://sing.stanford.edu/cs303-sp11/papers/1954-Fitts.pdf>.

Golomb, B. A., Koperski, S., and White, H. L., 2012. Association Between More Frequent Chocolate Consumption and Lower Body Mass Index. *Archives of Internal Medicine*, 172(6), p.519-521.

Gribbin, J., 1984. *In search of Schrodinger's cat*. London: Black Swan Books.

HERITAGE Family Study research consortium, 2012. *The HERITAGE Family Study*. Available at:<http://www.pbrc.edu/heritage/index.html>.

MacLeod, C. M., 1991. Half a century of research on the Stroop effect: an integrative review. *Psychological Bulletin*, 109(2) p.163–203. Available at: < http://arts.uwaterloo.ca/~cmacleod/Research/Articles/strooprev.pdf>.

Mynatt, C.R., Doherty, M.E., and Tweney, R.D., 1978. Consequences of confirmation and disconfirmation in a simulated research environment. *The Quarterly Journal of Experimental Psychology* 30.3: 395-406.

Rose, S., Bisson, J., & Wessely, S. (2003). A systematic review of single-session psychological interventions ('debriefing') following trauma. *Psychotherapy and psychosomatics*, *72*(4), 176-184. Available at: http://simonwessely.com/Downloads/Publications/Other_p/93.pdf

Schoemaker, P. J. (1993). Determinants of risk-taking: Behavioral and economic views. *Journal of Risk and Uncertainty*, *6*(1), 49-73. Available at: https://www.researchgate.net/profile/Paul_Schoemaker/publication/5152565_Determinants_of_Risk-Taking_Behavioral_and_Economic_Views/links/0c9605325c1413959a000000.pdf

Schunn, C. D., & Anderson, J. R. (1999). The generality/specificity of expertise in scientific reasoning. *Cognitive science*, *23*(3), 337-370. Available at: http://onlinelibrary.wiley.com/doi/10.1207/s15516709cog2303_3/pdf

Shepard, R. and Metzler. J., 1971. Mental rotation of three dimensional objects. *Science*, 171(972), p.701-3. Available at:<http://www.cs.virginia.edu/~evans/cs1120-f09/ps/ps3/mental-rotation.pdf >.

Sokal, A. (2010). *Beyond the hoax: Science, philosophy and culture*. OUP Oxford.

Stanovich, K. E. and West, R. F., 1998. Individual Differences in Rational Thought. *Journal of Experimental Psychology: General*, 127(2), p.161-188. Available at: <http://web.mac.com/kstanovich/Site/Research_on_Reasoning_files/Stanovich-Ind-Dif-JEPG98.pdf>.

Weber, E. U., Blais, A.-R., & Betz, N. (2002). A domain-specific risk-attitude scale: Measuring risk perceptions and risk behaviors. *Journal of Behavioral Decision Making*, 15, 263-290.

14. Appendix: Philosophy of science issues

This article focused on efficiency and on the characterization of phenomena, arguing that many explanations in science – perhaps all – try to account for one familiar phenomenon in terms of one or more others. This last idea is a theory of **scientific explanations**.

This position does quite well against the famous debates and problems in the philosophy of science.

The No Miracles Argument: This is the argument that the theories of science, even when they involve unobservable entities, must surely be true because otherwise it would be a miracle that they are so helpful in guiding technology.

The characterization of phenomena explains that this success could be because:

- characterizations of phenomena capture behaviour of reality that is inherently somewhat predictable;
- simple characterizations of phenomena telling us what to expect when we take particular actions can be useful even if we don't know why they work; and
- explanations of one phenomenon in terms of another still need to make predictions that (mostly) match repeatable phenomena, so they have predictive value even if they are not correct.

Pessimistic Meta Induction: Some past scientific theories have been successful at predictions for a while before being replaced by better theories. This is a reason for thinking that current theories might be wrong even though they seem to be doing well at the moment.

Characterizations of phenomena improve by finding better and better ways to analyse situations within which phenomena are observed and characterized. They also improve by becoming more accurate, and extending so that they work over wider ranges of circumstances, and with more variety of circumstances. Occasionally several phenomena thought to be separate and having their own explanations are explained in a new way by just one set of phenomena, giving a satisfying sense of unification. In these ways they become more effective and more efficient.

Under-determination of Theories by Evidence: While it is true that a variety of theories might account for the evidence we have, even in the worst case explanations using characterizations of phenomena may well provide useful predictive abilities even though they are only one explanation of many, at least within the range of situations where the usual behaviour is well known.

In the best case the explanation works much better and is able to make successful predictions of phenomena outside the range to which it was initially matched.

Unobservable Entities: Philosophers of science have argued a lot about the status of unobservable entities. However, 'unobservability' on its own is not really the issue. We can explain the extinction of dinosaurs in terms of familiar phenomena, such as living reptiles, even though we can never observe living dinosaurs. What really causes problems is trying to build explanations using phenomena that we have only imagined, rather than using phenomena we have observed and characterized.

Cognitive Relativism: One of the frequent arguments of groups that do not like science (e.g. 'postmodern' philosophers, some sociologists, some religious groups, some feminists, some people with other esoteric beliefs) is that science is no more than a social construct. This means that either the theories of science or the standards for what counts as good evidence are arbitrary choices made by scientists and that, in a different culture, those choices might have been made in other, equally valid ways. The argument then often continues that scientists have forced their world view on other cultures, other countries, women, and so on, which is not fair and not nice.

Sokal (2010) goes into this in painstaking detail, with many examples.

In reality not every research method is equally 'valid' because research methods differ in efficiency (in the sense described above). Some produce useful, reliable knowledge more efficiently than others.

For example, controlled, double-blind experiments are a strong way to root out technologies that don't work in situations where it is not obvious (e.g. new pharmaceuticals). In contrast, asking people if, in their experience and opinion, technologies work is not reliable and can lead to wasted resources and disappointment.

In many branches of science, research methods have become more stringent as people have learned from past mistakes. In particular, psychology and statistics during the late 20th Century and also during the 21st Century have revealed many ways that experimenters can be caught out.

Characterizations of phenomena are another argument against cognitive relativism. There's a reality out there (even if we don't know everything about it), and it's a reality that really does have some regularities in it, though we sometimes have to look hard to find them. Our characterizations of phenomena allow us some level of prediction and control, even when they are not perfect, and even when they may be replaced later by characterizations that are better structured and more economical.

In that sense they are true or at least close to the truth, regardless of whether people agree with them or agree with the methods used to create and justify them.

Explanations of phenomena in terms of characterizations of other phenomena can be true or at least close to the truth in the same way.

Falsification: The established debate is over whether science should focus on confirming theories or on falsifying them. Falsification implies making a deliberate effort to design studies that produce results that contradict currently interesting theories. In practice what happens is that researchers tend to produce results that support their theories but contradict those of their rivals.

Viewed in terms of characterising phenomena, things look different. Efficiently expanding the scope of a characterisation is probably best done by tweaking an existing study paradigm in a way that reveals new information but does not greatly upset what has been found already. It is rather like a miner following a rich vein of ore. This expansion is neither confirmation nor falsification. Efficiently closing in on the best explanation of a phenomenon in terms of other phenomena will involve competitively testing explanations against each other, using evidence. The evidence should be that which most efficiently reveals the leading explanations. In this competition there are winners and losers, so simultaneously there is falsification and its flipside, confirmation.