Critical Reasoning 11 - The Scientific Method

Alhazen (c. 965 - c. 1040 AD) *"Ibn al-Haytham" - Arabic Scholar, Polymath and Inventor of the Scientific Method in which he emphasised the need for experimental data and reproducibility of results.* (Wikipedia Scientific Method)

Science (from the Latin *scientia*, for "knowledge") is a way of knowing that distinguishes itself from other ways of knowing by being systematic, logical and which builds on the best available evidence and organises knowledge in the form of testable explanations and predictions about the world. Science is moreover inherently subject to correction and improvement upon discovery of better evidence. To that end, a good scientist ought to cultivate a healthy sceptical outlook as described by Alhazen, the inventor of the Scientific Method:

"The duty of the man who investigates the writings of scientists, if learning the truth is his goal, is to make himself an enemy of all that he reads, and, .. attack it from every side. He should also suspect himself as he performs his critical examination of it, so that he may avoid falling into either prejudice or leniency." (Alhazen)[1]

Science therefore differs from mathematical and logical knowledge by, in addition, requiring observational evidence and refinement. How is it then that Aristotle, the father of both Logic and Biology could have written in 350 BC:

> Males have more teeth than females in the case of men, sheep, goats, and swine; in the case of other animals observations have not yet been made: … (*History of Animals*, Bk 2 Ch. 3)

Clearly logic is not enough alone: either he didn't check or he didn't check very well. Compare this with following extract from the same work in which he describes the hectocotylus:

> The octopus, by the way, uses his feelers either as feet or hands; with the two which stand over his mouth he draws in food, and the last of his feelers he employs in the act of copulation; and this last one, by the way, is extremely sharp, is exceptional as being of a whitish colour, and at its extremity is bifurcate; that is to say, it has an additional something on the rachis, and by rachis is meant the smooth surface or edge of the arm on the far side from the suckers. (*History of Animals*, Bk 4. Ch. 1)

Clearly the latter could have come from a contemporary wildlife documentary; yet one of the problems of Aristotle's "scientific" descriptions here is lack of consistency. Scientific observations differ from every day observations in that they are both systematic and goal directed. That goal is the establishment of a well-established explanation of some aspect of the natural world, *i.e.* a

[1] Sabra, 2003

Scientific Theory (from the Greek θεωρέω (theoreo) for "I look at, view, consider, examine." (Wikipedia: Scientific theory)

Every scientific theory begins with a guess as a proposed explanation for some natural phenomenon or more formally an **hypothesis** (from the Greek ὑπόθεσις (hypothesis) for "that which is put under" or "supposition.") A good scientific hypothesis should be more than just a guess: it should be an educated guess based upon the best available knowledge at the time. More importantly it should be testable. *E.g.* the hypothesis: "The Universe just doubled in size last night," is a guess but it is not a *scientific* hypothesis because it is untestable. If it were true, then all the measuring instruments of length, such as rulers and yardsticks, would also have doubled in size last night and so would measure the same length today. If the hypothesis were false, on the other hand, they would also measure the same length today. Clearly, we would be none the wiser, so although such an hypothesis is meaningful (*pace* verificationism) it is untestable in practice and therefore unscientific. Furthermore, as we saw in Classic Text 09, an hypothesis, and by extension, a theory must also be falsifiable. This one is neither testable nor falsifiable.

Some hypotheses that relate to events that are "once off" occurrences, such as the Big Bang, the break up Gondwanaland, speciation within ancient lineages *etc.* are obviously not reproducible. However, hypotheses concerning them may be indirectly testable by the effects they leave behind, such as the red-shifted pattern of star light, the cosmic background radiation, geological clues, fossils and so on. Similarly, hypotheses that are too dangerous or unethical to test directly may be suitably tested indirectly.

The most basic type of scientific hypotheses runs along the lines of: "If I do x then y will happen," in which x and y are observable events, such as: "If I heat blue copper sulphate then it will turn grey." More frequently we want to frame an hypothesis in terms of one or more variables that are not directly observable such as "intelligence," "thirst," "kidney function..." *etc.* In such cases we want to do more than just observe instances of "intelligence," "thirst," "kidney function" and so on, we want to **quantify** them. *E.g.* Is that woman very intelligent? How intelligent? Are that man's kidneys functioning poorly? How poorly?

Operationalisation is the means by which we define the measurement of a phenomenon that is not directly measurable, as indicated by the existence of other phenomena that are directly measurable. In the examples above these are typically measured by proxy in the form of a psychometric test, a fluid or water deprivation test and creatinine clearance rate, respectively. Operationalisation can also be helpful in specifying the extension of a concept *i.e.* in describing what is and is not a part of that concept. *E.g.* we all have an idea of what we mean by "health." In medicine however, the concept of health might be operationalised by such proxies as body mass index, blood pressure or tobacco smoking. Indeed some phenomena that are very difficult or impossible to observe directly (*i.e.* **latent** phenomena) can only be inferred by observing their effects. Think of gravity for example. (Adapted from Wikipedia: Operationalization)

The **goal** of a good scientific hypothesis should be to explain the purpose and direction of research and to identify the variables involved. An hypothesis that is already widely accepted and forms the basis for further research is known as a **working hypothesis**, in which case the direction of research and the variables involved will already be known. For a good scientific hypothesis to be tested it must be stated precisely in the form of a proposition, either in natural language or in the form of equations, that can be verified and falsified. (yourdictionary.com: Examples of Hypothesis)

Hypothesis testing is seldom as simple as watching to see that y's follow x's as evidence for an hypothesis such as: "If I do x then y will happen." If you are dealing with complex systems, for example (and even molecules are complex,) then you might have great difficulty teasing out **confounding variables** that might give the appearance of a relation when in fact they are spurious. See: "Bananas & Fecundity" at right, which also illustrates how **correlation** need not imply **causation**. Also the mere act of measurement or observation may produce an **artifact** or error in the perception or representation of what is being studied. In the field of microscopy for example, an artifact may appear under the microscope depending on the way a slide was prepared. This might cause an inexperienced histologist to mistake it for something significant, whereas it is irrelevant. Very often however, relations between variables can only be discovered and quantified using **statistical hypothesis testing**, which is employed by researches from the Physical to the Life and Social Sciences. Basics of this methodology are introduced over several Critical Reasoning units, beginning with background knowledge on probability in Critical Reasoning 10 and 13.

Bananas & Fecundity (case study)

After WWII there was a corresponding increase in banana imports into Britten and the number of children women were baring on average. Was there a link? Were the increased number of bananas consumed having an effect on British woman so that they were able to bare more children? No! The confounding factor was the end of WWII: Ships that been previously assigned to the war effort were now free to resume trade internationally, hence the increased imports of bananas. On the other hand, women who had withheld baring children so as not to have them born into a time of war, now felt free to do so. Hence, the baby-boomer generation. Whatever correlation there was between such variables does not count as evidence for a causal relation.

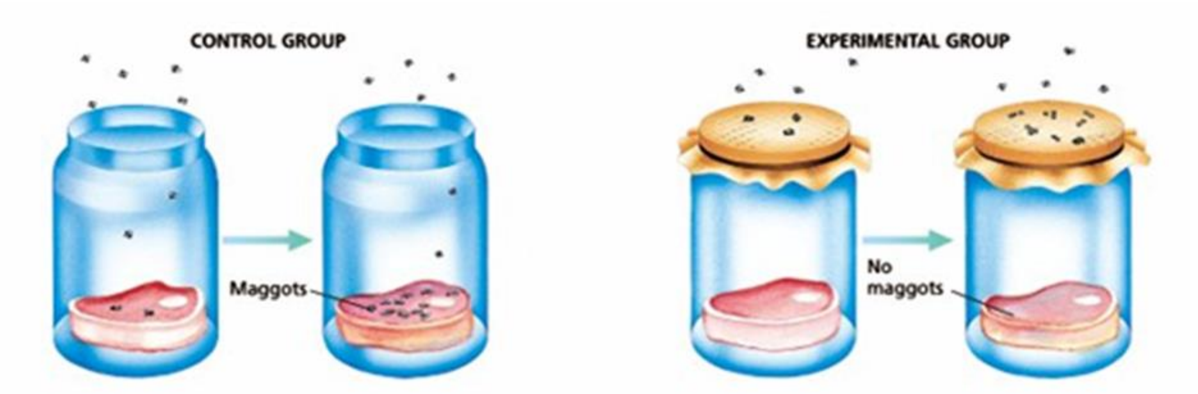
Besides close observation, the gold standard any scientific hypothesis testing remains experimentation. An **experiment** is an orderly procedure for establishing whether observations of the real world agree with or conflict with the predictions of an hypothesis. To the extent that they agree with such predictions they give weight to such an hypothesis. On the other hand, sometimes even a single contrary observation, such as seeing a black swan, can refute a long held "truism" such as "all swans are white." (See Classic Text 09)

A **controlled experiment** meanwhile, compares the outcome of one or more experimental samples with that of a control sample. Typically the experimental sample will be subject to some sort of action which influences the effect that is being tested (*i.e.* the independent variable.) The control sample on the other hand, will be practically identical to the experimental sample except for the action subject to the experimental sample. This procedure is necessary to isolate any outside factors that may be influencing the experiment. (Wikipedia: Experiment)

Spontaneous Generation (Case Study)

The simplicity and power of this method is illustrated by what was, arguably, the first controlled biological experiment by the Italian physician, naturalist, and poet Francesco Redi (1626 - 1697.) Published in 1668 as *Esperienze Intorno alla Generazione degl'Insetti* (Experiments on the Generation of Insects,) Redi sought to refute the Aristotelian notion of *abiogenesis* or "spontaneous generation." Until that time people believed that vermin such as mice, fleas and maggots arose spontaneously from such inanimate material as straw, dust and dead flesh respectively.

Redi performed a simple but decisive experiment to test the theory. He took six identical jars and divided them into two groups. In the first pair of jars, one from each group, he deposited an unknown object; in the second pair, a dead fish; and in the third, a raw piece of veal. He then covered the first group of jars with fine gauze that would admit only air. The second group of jars were left open. After several days maggots appeared on the objects in the jars in which flies had been able to land, but not the other. See diagram below.



In a second experiment, Redi placed meat in three jars. One was left open, another covered with the same gauze he had used in the previous experiment, while the third he closed with a cork (though some accounts mention parchment.) Again, maggots appeared on the meat in the open jar, however they also appeared on the gauze of the second jar but did not survive. No maggots appeared in the jar stopped with a cork. (Wikipedia: Francesco Redi)

According to the theory of spontaneous generation, maggots arise spontaneously from decaying meat. By allowing air to pass freely into both groups of jars in the first experiment Redi insured that their contents would decay, however by preventing flies from visiting the contents of the jars in the experimental group he was able to eliminate decay itself as the cause of the appearance of the maggots.

Redi's hypothesis, that both pieces of meat would decay, but that only the one that had physical contact with the flies would produce maggots, meets all the criteria a good hypothesis in guiding research. It:

- Is practically testable, logical and uses precise language.
- States the purpose of the research: To test the theory of spontaneous generation.
- Identifies what variables are to be used: Covered and uncovered jars with 3 different contents.

Redi's controlled experiment meanwhile is both simple and rigorous. The **controlled variables** are:

- Uniform jars
- Type of sample: unknown, fish or veal

- Location
- Temperature
- Time (duration)

The **independent** or manipulated **variable** was the gauze covering preventing the flies from coming into physical contact with the sample. The **dependent** or responding **variable**, on the other hand was whether or not maggots would appear on the sample. (See box right.)

Independent vs. Dependent and Extraneous or Controlled Variables

In the experimental context the **independent** or **input variable** represents the variable that the experimenter manipulates in order ascertain whether it is the cause of some effect. The **dependent** or **output variable**, on the other hand, literally depends on the independent variable in some way and is observed to see whether it is the effect of the change made to the independent variable.

An **extraneous** or **controlled variable** may or may not alter either the independent or dependent variables in some way. However because it is not really the focus the research proper it must be controlled or kept constant so as not to interfere with the outcome of the experiment.

Above all Redi's experiment is **reproducible**. Anyone with an interest in the question is free to repeat the experiment for himself. Un-published or secretive experiments or those that cannot be reproduced do not form part of Science. At considerable risk to himself, Redi actually managed to publish his work, even though it contradicted Aristotle on the matter. Galileo Galilei before him, recall had not been so fortunate, having been found "vehemently suspect of heresy" and sentenced to house arrest for contradicting Scripture, which accorded with the Aristotelian geocentric view of the solar system.

Once an hypothesis has been widely accepted it attains the status of a **law**. Indeed Redi was so confident of his conclusion that he generalised it to all life according to the law *omne vivum ex vivo* "All life (comes) from life." Unlike the laws of mathematics and of logic such as the law of cosines or the law of non-contradiction, which are certain, scientific laws remain refutable and subject to refinement. Should scientists create truly synthetic life *de novo* in the laboratory, and there is every expectation that this will be achieved soon, then Redi's law will have to admit of an exception.

A **Scientific Theory** represents a set of relationships among concepts, supported by group of scientific hypotheses that explain some aspect of the natural world. Such a theoretical structure is more readily visualised as a web of relations, with anchoring stands providing scaffolding and others providing collateral support. Another structural commonality is that they should all "hang together." Like organic webs moreover, scientific theories are dynamic and evolving rather than static or monolithic. (See the Scientific Method below.)

Unlike "pet" theories or other unscientific theories, scientific theories are well-substantiated through repeated observation and experimentation. Ideally, all scientific theories should:

1. make empirically falsifiable predictions consistent with such a theory.
2. be well supported by independent strands of evidence, rather than a single observation.
3. be consistent with, and at least as accurate as, pre-existing findings.

4. be flexible enough to accommodate minor adaptations to account for new observations that do not quite fit it perfectly, thus increasing its predictive capability over time.
5. be **parsimonious** in its explanatory power and economical in the entities it postulates. (See Occam's razor, at right.) (Criteria adapted from Wikipedia: Scientific theory)
6. Although not a diagnostic criterion, the greatest of Scientific Theories have the quality of elegance about the,

The need for a scientific theory to be falsifiable, over and above merely verifiable, was discussed in Classic Text 09. Ironically a scientific theory's strength lies in its *vulnerability* to falsification. Indeed it *requires* that conditions that might render it false be tested and, through repeatedly failing to be falsified, gains credibility. That is not to say that a scientific theory gains in "truth" in this way because at any time it may yet be falsified. Instead, so long as there is evidence to support it and it is not being falsified, it is provisionally accepted a true *representation* of some aspect of the world.

Ockham's razor or *lex parsimoniae* in Latin, named after its inventor, the English Theologian and Philosopher, William of Ockham (c. 1287-1347,) is an heuristic device that states literally, "entities must not be multiplied beyond necessity." (From the Latin: "*entia non multiplicanda praeter necessitate.*") In other words, among competing hypotheses, the one with the fewest assumptions should be favoured. In practice this means that often the simplest hypothesis is preferred although this is not the intended meaning. According to Ockham's razor, a complex hypothesis with very few assumptions should, for example, be favoured over a simpler one that makes many assumptions.

Of course there are theories such as the Theory of Universal Gravitation and the Germ Theory of Disease to which most people have grown so accustomed that they incorrectly regarded them as proven, rather than provisional. In the case of gravitation for example, Newton's Law of Universal Gravitation is now regarded as only a limiting case of Einstein's General Theory of Relativity, which in turn has not yet been reconciled with Quantum Field Theory or Quantum Mechanics. In the case of germ theory, on the other hand, it has been found that a multitude of diseases are caused, not by germs, but by environmental factors or even the genetic code within our every cell.

The second criterion above, that scientific theories be well supported by independent strands of evidence, rather than a single observation, is required to rule out coincidental or anomalous regularities within only one or a very few contexts. The chance "discovery," for example of a single "cure" for nosebleeds in one individual at one time only is really no discovery at all, even less so if it is unsupported by any other threads of evidence. To return to the web analogy: in the same way that a web, or indeed a tent, must be supported by multiple anchoring threads or guy-ropes respectively, so a scientific theory must be supported by multiple strands of evidence from independent fields. As noted above, theories of events or occurrences that are "once off" can still be tested from via indirect lines of evidence. The bubonic plague or "Black Death" of 14th Century Europe, for example, is one such irreproducible event. Far from being discouraged, scientists have studied the plague and proposed *post hoc* falsifiable hypotheses and **retrodictions** ("predictions" about the past.) These have been drawn from such diverse fields as Microbiology, Entomology and

Sociology, so that now we have an unparalleled scientific understanding of one of the darkest hours in human history.

The third criterion, that scientific theories should be consistent with, and at least as accurate as, pre-existing findings is a no-brainer. Any theory that is less accurate than its predecessor or that does not accommodate current or pre-existing finding, is a step backwards. If anything, an improved scientific theory should encompass and explain more, rather than fewer findings.

These first three criteria are essential. It is hard to imagine any theory, worthy of the name "Scientific," that is either unverifiable or irrefutable or that is sustained by a single observation or that is less accurate or less consistent with fewer findings than before. The following criteria (4 - 6) are highly characteristic of scientific theories, but they are not, on their own, diagnostic of them.

When a new piece of evidence is acquired that does not quite fit an established scientific theory, there are three possibilities: discard the discrepancies in evidence, discard the theory wholesale or make minor adjustments to the theory so as to accommodate the discrepancies. The first is dishonest, the second extreme, while the third is somewhat mercurial. Concerning the latter: the virtue in making incremental adjustments to an established theory from time to time, as new but slightly divergent discoveries are made, is that the predictive power of the theory may be ratcheted up with each successive round of near, but not quite perfect, observations. The drawback, however is that one may be started down the path to a "**death by a thousand qualifications**" discussed in Critical Reasoning 06. Recall that, if one is willing to qualify a theory, an idea or a concept at every turn, it will surely end in the "death" of that theory, idea, or concept. Like people, theories, ideas and concepts require boundaries that define them, *i.e.* between what they are and are not. A slight adjustment is unlikely to make much difference, however if one steadily chips away at a boundary or condition, the cumulative result is likely to be the end of it.

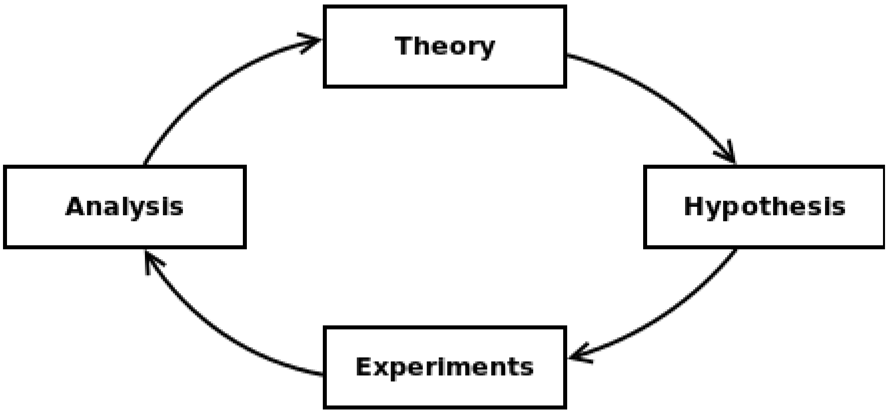
Parsimony or economy, for which there are no objective measures, are characteristic of the greatest scientific theories to date. A Great Scientific Theory is one that explains a great deal about many aspects of the world with the minimum of assumptions. According to Isaac Newton in his *Principia*, "We are to admit no more causes of natural things than such as are both true and sufficient to explain their appearances. Therefore, to the same natural effects we must, so far as possible, assign the same causes." While Newton called these "philosophising rules," they would be classed as heuristics today. (*Principia*, 1726 edition.)

A Theory of Everything that fully explains all physical aspects of the Universe from, ideally, only a handful of axioms or assumptions, modelled by one or only a handful of interlinked equations, would also be supremely parsimonious or economical. We do not know today whether we shall ever arrive at such theory. Perhaps reality will prove to be fundamentally and irreducibly complex, in which case we ought not to be guided too zealously by such considerations. On the other hand not all Sciences are like Theoretical Physics and no two Sciences from Virology to Psychology are driven by the need for parsimony or economy to the same degree.

The quality of elegance or aesthetics resides not in a theory itself but in its appreciation. Humans have an insatiable need to understand the world around them and to structure their knowledge. The greatest scientific theories address both of these. To be in awe of the majesty of nature is a powerful emotion, however to be in awe of a Universe that is humanly intelligible, one that unfolds according to discoverable laws and regularities is an altogether different order of aesthetic.

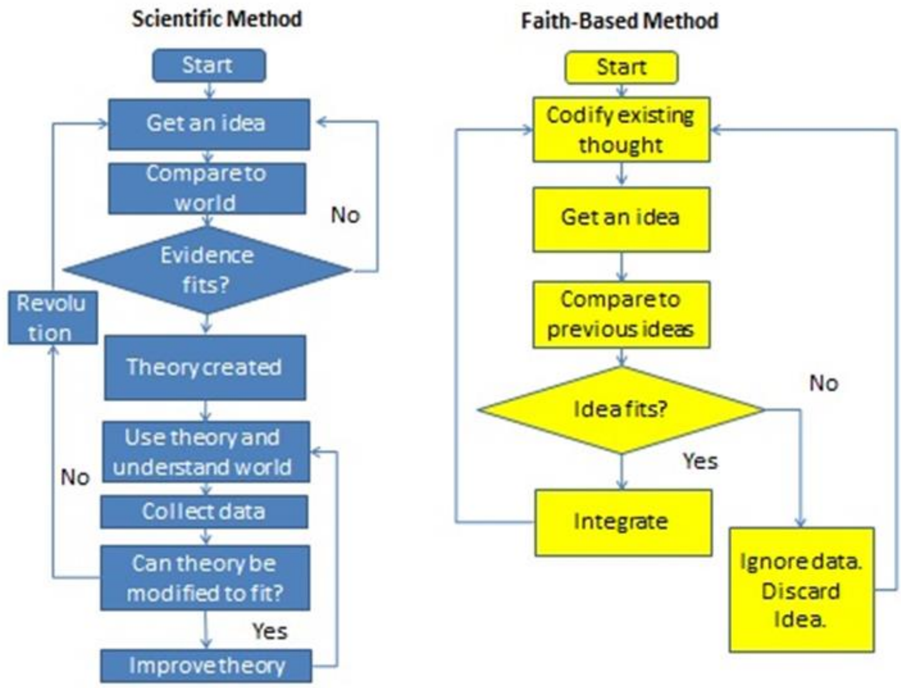
The Scientific Method

In discussing scientific theories in the form of a linear narrative, as above, it is possible to imagine them as ornaments or curious objects on display, that are taken down from time to time to be tweaked, polished or talked about by the cognoscenti. On the contrary, scientific theories are dynamic, almost organic entities: always in play, ever evolving, giving rise to other such theories, informing yet others, occasionally becoming extinct (and rarely, very rarely resurrected.) In the following diagram the scientific method is represented as a continuous process.



In an ideal world one would begin with an hypothesis, devise experiments to test the hypothesis and analyse the results. As the number of confirmed (or at least not refuted) hypotheses accrues, these suggest the outlines of a theory, which in turn generates further hypotheses for experimental testing, and analysis. With each turn of the cycle the theory in question would (in theory) come more closely to approximate of some aspect of the world being studied. Of course the real world is more complex, with historical, logistical, ideological, ethical and personal considerations. However what this diagram does depict is an ongoing, cumulative process, quite unlike any of our innate cognitive processes by which we navigate the world of ideas from day to day.

Before the advent of the scientific method, including the early flourishings of ancient Greek and Chinese Philosophy, men (and it was almost exclusively men,) reasoned according to Faith, a method, which still persists alongside the former, sometimes even within the same head. Both methods are presented side-by-side below in the form of linearized flowcharts. What makes the faith based system so very different from the scientific method is the complete absence of reality testing, the irrefutability of previous ideas and the inability of old ideas to evolve into radically new ones, depending on the context. One process moreover, that of Scientific Revolutions (extreme left) which from time to time upend and overthrow established paradigms, in favour of more substantive and productive world views, are a complete anathema to Faith. And although some argue that there is no conflict between Science and Faith, not even a faith practitioner would prefer the safety or quality of his medicines or drinking water to be guaranteed by the Faith-Based Method. Below we discuss the universally acknowledged method of randomised controlled trials as instance of scientific experimentation by which to assess just such examples.



A schematic representation of the Scientific and Faith-Based methods. Posted on December 20, 2012 by Paul Christensen *http://www.intlcorn.com/seedsiteblog/?p=1065*

Scientific Modelling

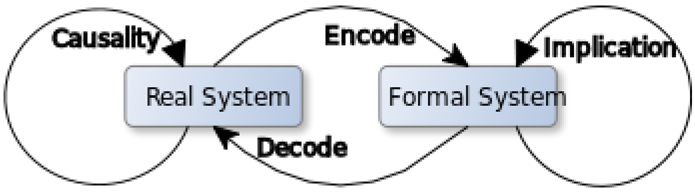
Because certain phenomena are extremely complex, such as the functioning of the human immune system, the global climate, the rapid progress of an epidemic and so on, it impractical to account for or impossible to simultaneously understand all the features involved. Instead, scientists and increasingly some philosophers, try to identify relevant aspects of real world phenomenon and compare them with a conceptual representation of such phenomena. In the same way that an architect might build and test a model of a building he intends to design, so a scientific model may be created or discovered that addresses certain aims:

- A **conceptual model** is one by which to better understand phenomena.
- An **operational model** meanwhile, helps to operationalise relevant variables.
- A **mathematical model** relates and quantifies such variables, and
- **Graphical models** are used in order to visualize the subject. (Wikipedia: Scientific Modelling)

Today, with the aid of computers, many scientific models are embodied in the form of software. Although a mathematical or statistical model may consist of a few (or even many) combined equations, these will typically be encoded in spreadsheet into which raw data can be entered. Implementing a model in such, or indeed any other way, is known as a **simulation**. One great advantage in employing a model is that theoretically, there is no upper limit on how many simulations may be run.

***In* vivo** models, such as laboratory mice, are commonly used in Physiology or Medicine because they resemble humans in some salient respect. This allows researchers to infer processes that are likely to occur in humans from those known to occur in a non-human animal model. Apart from the ethically dubious practice of harming or killing other conscious animals for human benefit, different species are not isomorphic in all respects. *E.g.* A drug that successfully reduces cholesterol in rats may cause unbearable headaches in humans. After all, the drug may have produced similar headaches in rats, only they weren't able to tell us about them.

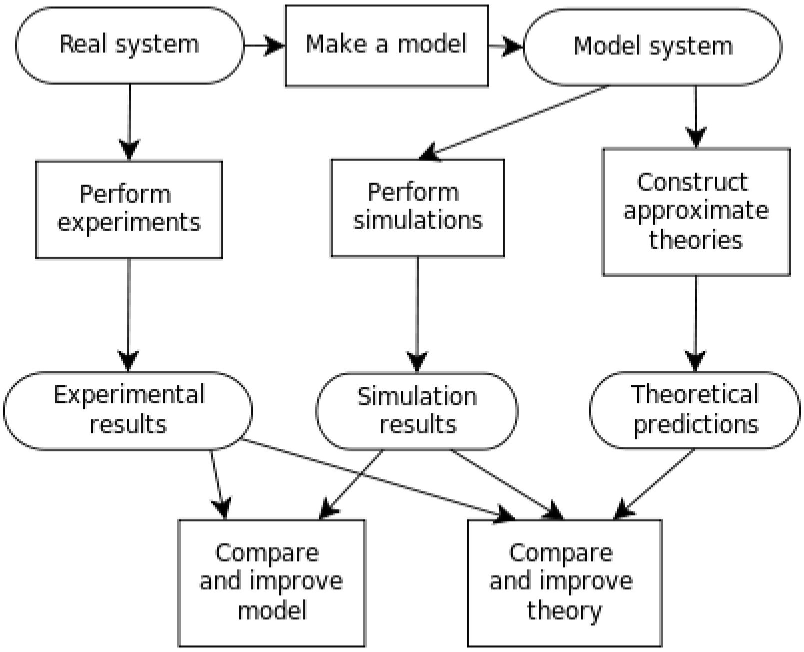
In vitro or "in glass" models rely on investigations with cultures of human or other tissue in Petri dishes or other laboratory ware. This circumvents any ethical concerns, so long as the samples are obtained with consent. Another great benefit of *in vitro* models is that, so long as researchers can coax the tissue to multiply and grow, they have an endless supply of living material on which to run simulations of *in vivo* processes. The major detraction however is that not all tissues behave the same way in laboratory ware as they do in the context of the living body, in which case *in vivo* studies may have to be carried out anyway.



The process of building and interpreting a model can be understood with the aid of the following, deceptively simple, unaccredited diagram, at right, from Wikipedia: Scientific Modelling. On the left is a Real System with causal relations that we are trying to understand. On the right is a Formal System which is a representation of the real system on the left. Formal systems are constituted by their internal relations and so, when the behaviour of one component follows another, one will be able to draw formal implications within the system.

Although the performance of the Formal System replicates some aspect of the Real System, they are actually unrelated: What happens causally in the Real System has no logical bearing on the Formal System. For this reason, information about the Real System must be encoded and fed into the Formal System. On the other hand, what implications arise within the Formal System must be decoded and compared to the Real System. Scientific models are regarded as "true" to the extent that predictions or statements drawn from the Formal System map or track the real world. (Wikipedia: Scientific Modelling)

Simulations can therefore be thought of as virtual experiments, the outcome of which can be fed back into either the same model or theory in order to improve either or both. In the diagram at right the processes of experimentation, simulation and theoretical construction are depicted as running in parallel, though in practice, they need not.



A Schematic Representation of Model Building and the Relations between Experiment, Simulation, and Theory. (Wikipedia: Computer Simulation)

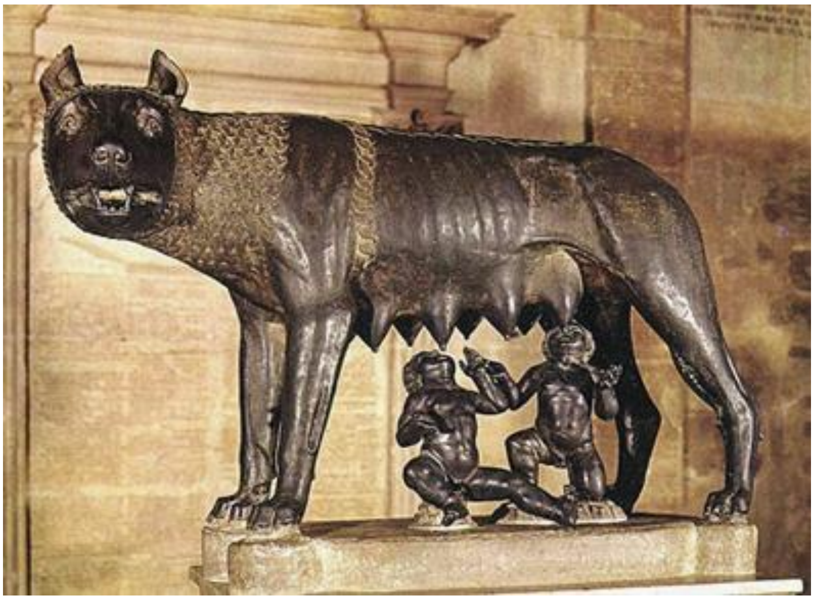
Natural Experiments

A natural experiment is one in which Nature determines which individuals are exposed to experimental or control conditions. Researchers have no control over natural experiments in either the scientific or vernacular sense of the word “control.” They are unbidden and their results can only be observed after the fact, sometimes in awe, sometime in horror.

The Unspeakable Experiment is one such example. Linguists, Psychologists, Sociologists and Anthropologists have long wondered about the contribution of language to human nature. Some have argued that it is fundamental to what it means to be human, others that it plays more of an ancillary role. The only way to find out for sure would be to perform an experiment, but only a monster would do so. In the unspeakable experiment, one hundred newly born babies would be removed from their mothers at birth and randomly divided into two groups. The experimental group would be prevented from hearing speech or seeing lip movements or signing. The control group would be allowed to develop normally. When the subjects reach adulthood they would be subject to a battery of linguistic and psychometric tests that would reveal the true role of language in their development. The *double entendre* here is that the experimental group would literally be rendered “unspeakable,” while the experimental design and its execution, should it ever be carried out, would itself be an “unspeakable evil” and worth of the harshest punishment.

The existence of **feral children**, who have been either confined by their abusive parents or abandoned and raised by other animals such as wolves, represent one of Nature’s most horrific “experiments.” Without performing the unspeakable experiment we now know, from feral children, that without language (as well as human care and social behaviour,) human psychological development is severely compromised and that, beyond a critical window period, none ever learns to speak or socialise normally.

Feral children are very much the stuff legend but there is a growing list of well documented cases. See “Wikipedia: Feral Child” for a start, including external links.



The Stuff of Legends: Romulus and Remus, the Founders of Rome, Suckled by the Capitoline She-Wolf.

Double Blind Randomized Controlled Trials

A **randomised controlled trial** (RCT) is a controlled experiment for assessing the efficacy or effectiveness of some clinical intervention, be it pharmaceutical, surgical or other therapeutic. A RCT is so called because participants are randomly assigned to one of two or more groups, at least one of which is the experimental group and another, the control group. The experimental group is given the new treatment, while the control group is given a sham treatment or placebo. (See Critical Reasoning 6: Placebo *vs. Nocebo* box.)

Ideally, a RCT should be double blinded, *i.e.* neither participants, nor the caregivers should know whether they are receiving or dispensing the experimental or sham treatment. To this end, both the experimental and sham treatments should resemble each other in every respect other than the supposedly active substance or activity being tested. In other words if a pill or capsule containing an active substance is being tested, the placebo and treatment should look, taste smell *etc.* indistinguishable from each other, so that not even a caregiver can tell one apart from the other. Of course the person or agency running the experiment will know one from another because they will be coded in some way such as numerically or by use of bar codes.

The double blinding process is required because both those receiving and those dispensing a treatment may act in subtle (and sometimes not so subtlety) different ways, if they know just what they are receiving or dispensing, respectively. *E.g.* a doctor who knows that she is giving the active substance to a participant may convey an enthusiastic sentiment via her body language and facial expression alone. This might cause the participant to feel subconsciously better (or worse,) about the treatment he is receiving than if he didn't know one way or the other, thus introducing a confounding variable. On the other hand, sometimes it is impossible to either doubly or singly blind a trial. In the case of experimental treatments offered by physiotherapists, both the therapist and participant know if, for example, a particular massage to the shoulder is being applied: the participant knows what he is feeling and the therapist knows what he is doing.

Ethical considerations are also paramount. One cannot, for example, divide an elderly group of men who are suffering from heart disease into two groups, one in which a new surgical procedure is carried out on the heart and another in which a sham surgical procedure is carried out on the heart. One can never endanger the wellbeing of participants in order to advance the interests of a trial. "First, do no harm" is the ancient maxim expressing one of the principal precepts of bioethics but it ought equally to be adopted by all disciplines, with the notion of harm extended to include harm to future generations and the environment. Bioethics is a discipline in itself but what must be emphasised here is that an ethical frameworks should be established to which a trial should be tailored (not the other way round.)

Double blind RCT are regarded as the gold standard of clinical trials and are required in most contraries before a clinical procedure or a drug is registered for use. However, because they require a large number of participants to deliver statistically significant results and because they must be compensated for time lost and cost of transport as well as followed up or monitored after such trials, they are extremely expensive. Also pharmaceutical corporations tend to selectively report only those trials in which their prospective products fare well, ignoring often, the majority of trials in which no benefit could be demonstrated.

Indeed, in his book “Bad Pharma: How Drug Companies Mislead Doctors and Harm Patients” (right,) the British physician and academic Ben Goldacre has documented a litany of malpractice, unethical behaviour and downright dishonesty by some of the most well-known and trusted pharmaceutical companies in both the U.K. and U.S.A. (many of which have subsidiaries in South Africa.) Goldacre is at pains to remind the reader, several times over, that because RCT trials are comparing an active substance to a placebo containing nothing (of substance,) a prospective drug need only perform *better than nothing* to be considered for registration. This is setting the bar as low as possible for it to be set. In the case that a prospective drug is compared to an existing drug, it is often tried against the worst performing drug within its class. In such a case the standard of evidence is set a depressingly low: *better than worst*.

Instead of despairing, Goldacre proposes numerous positive measures by which Big Pharma can be better regulated, more accountable and transparent, including publishing of all trial data that does not compromise patient privilege. Rather than rehearse Goldacre’s arguments here, anyone who is dependent on medication or who has a loved one who is, should read this book, as should anyone interested in the nature of evidence gathering by corporations. In conclusion, the philosophical lesson to be draw from Goldacre’s review should be that, although the Scientific Method strives to be open and objective, it is often subverted by corporate greed and less often for personal gain. It is, after all, a very human endeavour.

Task

Conduct an experiment with a view to demonstrating how the Scientific Method works in your own small way. The emphasis should be on small, or rather small and practicable.

Warning, do not attempt to conduct your own clinical drug trial either on yourself or others. You may land up in hospital or prison, or worse: a prison hospital. Even common drugs like aspirin can be very dangerous to deadly if misused.

Begin with a modest observation in need of an explanation. “We are here, but why?” is not suitable for now but perhaps, “I suspect my partner’s occasional sleepless nights have something to do what he drinks before bed, or possibly the moon, or both,” is certainly amenable to home experimentation.

Feedback

You may wish to judge the design and execution of your experiment against the following fictitious little one.

We wrote down our observation(s) together with our hypotheses. We made a list of observations **O_1** - **O_2** and hypotheses **H_1** - **H_3**.

O_1: My partner believes that the Moon disturbs his sleep: "…*deur die maan gepla.*" - his words, not mine.

O_2 I believe he sleeps more soundly when he has fewer caffeinated beverages before bedtime.

H_1: My partner's poor sleep is caused by more than usual caffeine intake before bed time.

H_2: My partner's poor sleep is caused by a brighter Moon.

H_3. My partner's poor sleep is caused by a combination of more than usual caffeine intake before bed time as well as a brighter Moon.

[By convention **H_1** - **H_n** are known as **alternative hypotheses**, but alternative to what? They are alternative to the default position *i.e.* a statement of what would be the case if the alternative hypotheses were not true. This is known as the null hypothesis (**H_0**). In this case:]

H_0 My partner's poor sleep is not caused by either more than usual caffeine intake before bedtime nor by an unusually bright moon.

[In Science, we assume that the null hypothesis is true, unless it can be shown otherwise. That was the purpose of this experiment.]

Next we needed to identify our variables and operationalise them. Up to now we have been very sloppy with our language *e.g.* "My partner's poor sleep… more than usual caffeine intake before bed time *etc.*" How poor is "poor?" How much is "more than usual caffeine intake" and how early "before bedtime" is early or rather late? Here is a list of variables involved in our experiment, together with the way in which we operationalised them.

V_1 Poor sleep: I asked my partner to fill in on a calendar each morning how he felt he slept the night before:

1 - very poorly
2 - poorly
3 - normal sleep
4 - good sleep
5 - very good sleep

V_2 Moonlight: Fortunately, the phases of the moon are clearly marked on the same calendar, so we scored the amount of moonlight as:

1 - New Moon
2 - Quarter Moon
3 - Half Moon
4 - Three Quarter Moon
5 - Full Moon

We did not bother about clear or cloudy skies because we believed that they would be just so much **noise** (unexplained variation) which would cancel out over the duration of the experiment.

V_3 Time before bedtime: My partner always goes to sleep after listening to the 11pm news on the radio in bed, so there was no difficulty in operationalising bedtime *i.e.* 11:05pm.

I also wanted to make sure that if we counted a caffeinated beverage *before* bedtime, then most of the caffeine in them would still be in his body at bedtime, so I looked up the **half-life** of caffeine (how much time it takes your body to get rid of half of it.) It is 5 - 7 hours and since we can't afford lab tests, we just picked the middle number *i.e.* 6 hours. Therefore, we decided to count all caffeinated beverages consumed 6 hours or less before bedtime *i.e.* 5pm or later, but none before 5pm.

V_4 Caffeinated beverages: Because not all caffeinated beverages are equal I looked up the ones my partner drinks and gave them a score according to how much is in each:

1 - Green or black tea
2 - Single coffee, espresso or cola
3 - Large coffee
4 - Double espresso
5 - Energy drinks with caffeine *and* other stimulants

<u>Method</u>: At 5pm every day I cleared the clutter around the sink and asked my partner to put all his empty caffeinated drinks he consumed after 5pm there so I could count them. I asked him to use a new mug for each beverage, which he does anyway. Since I am the first one up in the morning, I counted and scored all the caffeinated beverages he consumed the night before and marked the score for the previous night on the same calendar in a different colour, as well as the phase of the moon for that week. The trial ran for 4 months or 120 days, which provided plenty of data.

Our calendar was essentially our means of recording data and analysing it. Although we both know of statistical techniques that we could have used to analyse our data, we didn't know which one was appropriate for our study. Instead, we cut up and pasted together the four months of the calendar into seven-day strips: one strip for each phase of the moon. These we laid out on the floor side-by-side so that we had a strip of 120 squares long. We then took turns to independently note what trends we could observe and afterwards compared notes.

<u>Results and Discussion</u>: Our observations were in agreement: My partner was sleeping most poorly around Full Moon and best around New Moon. So clearly, this is not something in his imagination. Paradoxically, he was also drinking more caffeinated beverages after 5pm around Full Moon and fewest around New Moon. Although this suggests that hypothesis 3 is correct *i.e.* his poor sleep is caused by a combination of more than usual caffeine intake before bed time as well as a brighter moon, we cannot accept that it is the whole story. He and I are both above intelligent adults and know that caffeine before bedtime is the enemy of sleep. So arguably, something around Full Moon is causing him drink more caffeine around bedtime, in spite of his better judgement.

<u>Conclusion</u>: I believe the following is the correct scenario: Something, it may just be the recollection of previous sleepless nights, is making my partner anxious around Full Moon, which is causing him to loose sleep. The next day he feels exhausted and so resorts to pick-me-ups to counter the fatigue. This results in even worse sleep the same night, fuelled by caffeine. This cycle reinforces itself for several days until either the anxiety abates or he stops drinking so much caffeine, or preferably both.

The variable “caffeinated beverages,” which I was most careful to measure, now turns out to a confounding or nuisance variable which we are going to have to eliminate or control, should we take this experiment forward. However, persuading my partner to either forgo caffeine after 5pm or consume only a fixed amount and no more, every day for another 120 days will prove challenging. Also we are going to have to find a suitable instrument by which to regularly quantify “level of anxiety.” Those available on the internet are either unattested or very expensive and lengthy, which is quite unsuitable to our task. Although, our mutual friend Amy, who is registered for a module in Psychometry in her Honours years, has promised to help next time round.

I believe our experiment does demonstrate the Scientific Method in action, in however small a way. We have defined our variables, stated our hypotheses, and tested them against reality by experimentation rather than seeking opinion or faith in our own conviction. As it turns out, our results did not suggest a definitive answer. We will therefore have to modify our hypotheses and repeat our experiment, possibly more than once, and so come closer to the truth.

I am unconcerned by criticism that my experiment had only one participant because I set out to do a <u>little</u> experiment and in so doing, help my partner.

Kim

The Unsuitability of Pearson’s r for Kim’s Experiment

If you already have a smattering of statistical knowledge you may be wondering why Kim did not simply calculate Pearson’s correlation coefficient r using a spread sheet or a scientific calculator. The problem lies with the nature of Kim’s data. They are discontinuous and not monotonic. *I.e.* the go up in jumps from one integer to the next from 1 to 2 to 3 *etc.* (fractions of a cup are not counted) and they do not always rise or always fall. At times they rise and other times they fall over the entire interval.

Although we will describe a suitable test for quantifying the correlations among Kim’s data towards the end of the study units concerning statistics, the tendency to rush into a calculation just because we can, should be accompanied by a cautionary lesson. Modern computers and the more sophisticated scientific calculators are equipped to handle statistical calculations that in the past would have taken days by hand; however they are silent on the underlying assumptions required for each test. Pearson’s r for example is only equipped to handle data that are continuous and monotonic, failing which it will yield a distorted or even meaningless result. Kim was therefore right to have held back on a statistical test calculation before seeking assistance.

Critical Reasoning 13 deals with probability distributions and represents the first in a series of alternating study units dealing with statistical topics written for the non-specialist.

References:

GOLDACRE, B. (2012) *Bad Pharma How Drug Companies Mislead Doctors and Harm Patients*. : Fourth Estate Ltd. U.K. (Paperback: *Bad Pharma How Medicine is Broken, and How We Can Fix It.*)

SABRA, A. I. (2003) Ibn al-Haytham: Brief life of an Arab mathematician. *Harvard Magazine* (October - December)

WUDKA. J.(1998) The scientific method. http://physics.ucr.edu/~wudka/Physics7/Notes_www/node5.html