Sci Eng Ethics
DOI 10.1007/s11948-015-9635-3

ORIGINAL PAPER

Experiments: Why and How?

Sven Ove Hansson

Received: 3 September 2014 / Accepted: 14 February 2015

Abstract An experiment, in the standard scientific sense of the term, is a procedure in which some object of study is subjected to interventions (manipulations) that aim at obtaining a predictable outcome or at least predictable aspects of the outcome. The distinction between an experiment and a non-experimental observation is important since they are tailored to different epistemic needs. Experimentation has its origin in pre-scientific technological experiments that were undertaken in order to find the best technological means to achieve chosen ends. Important parts of the methodological arsenal of modern experimental science can be traced back to this pre-scientific, technological tradition. It is claimed that experimentation involves a unique combination of acting and observing, a combination whose unique epistemological properties have not yet been fully clarified.

Keywords Action-guiding · Experiment · Observation · Bias · Blinding · Randomization

Introduction

It is commonplace that we can distinguish between two types of knowledge. One of these can be called factual knowledge. It consists in being able to give an account of how things are. You prove your knowledge of facts by making the right statements. The other type of knowledge can be called action knowledge. You prove your action knowledge by performing actions that lead to a desired result. The distinction between factual knowledge and action knowledge is often associated with Ryle's

S. O. Hansson (✉)
Department of Philosophy and History, Royal Institute of Technology (KTH), Brinellvägen 32, 100 44 Stockholm, Sweden
e-mail: soh@kth.se

Published online: 28 February 2015

(1949, 1971 [1946]) account of knowing that and knowing how, but the distinction is of course much older. Leaving out some of the historical detail, it corresponds to Aristotle's distinction between episteme and techne.

This chapter will be concerned with scientific knowledge, or more precisely with the knowledge claims of science. (I use the term "science" in a wide sense including the humanities.)[1] It is important to recognize that science includes both types of knowledge. A large part of the science that is conducted in practice is devoted to action knowledge (techne); that is for instance the type of knowledge primarily promoted in technological, medical, veterinary, and agricultural faculties, and a large part of the knowledge developed in departments of economics, environmental science, and education research, to mention just a few examples.

For knowledge of fact (episteme) the central practical pursuit is *observation*. It is by observing the world that the astronomer, the archaeologist, and the geologist obtain the knowledge that proves their title to expertise in the respective areas. The central practical pursuit for action knowledge is of course *action*. It is by preventing and curing disease, not by just knowing the courses of diseases, that the physician proves her expertise[2], and something similar can be said about the knowledge of veterinarians and engineers.

But action and observation are not exclusive categories. In experiments the two are combined. And experiments are performed to obtain both knowledge of facts and action knowledge. The particle physicist who accelerates particles in order to observe their interaction at high energies does so in order to obtain factual knowledge; the clinical scientist who performs a treatment experiment (clinical trial) in order to learn how best to treat a particular disease is pursuing action knowledge. Experiments seem to straddle across the demarcation line between the two types of knowledge.

In my view, the combination of action and observation that takes place in experiments has specific epistemic features that make it important to distinguish experiments from observations that lack the active component. It is the purpose of this essay to incarnate this rather skeletal statement and hopefully shed some light on the nature of experimental activities.

What is an Experiment?

The word "experiment" is often used in an everyday sense that has very little to do with science. When you try a new recipe at the stove, you may call it "just an experiment" without implying any scientific systematicity. This usage seems to be close to the original meaning of the word. It has the same origin as "experience", namely the Latin verb "experiri" that means to try or put to test. We should keep this in mind when reading early philosophical texts. Roger Grossetest (1175–1253)

[1] This corresponds to the word "Wissenschaft" in German with its close analogues in some other Germanic languages. See Hansson (2013) for an argument why this wider disciplinary delimitation is more adequate than the traditional one in the English language.

[2] The therapeutic nihilists thought otherwise, see Wiesemann (1991).

had a large part in introducing the term "experiment" in natural philosophy, but he used the term both for active procedures and for the passive collection of experiences. (Eastwood 1968, 321; McEvoy 1982, 207–208) Francis Bacon (1561–1626) employed the word in the same broad sense (Klein 1996, 293; Pesic 1999; Klein 2005), and some of the "experiments" referred to in David Hume's (1711−1776) *Treatise* consist in the systematic recording of everyday experiences, without any element of action (Robison 2008).

As far as I can see it was in the first half of the nineteenth century that the modern meaning of the term "experiment" became established. It seems to have been promoted as a key component in attempts to account systematically and philosophically for the methods used in a long series of successful laboratory investigations in the physical and chemical sciences. In his influential *A preliminary discourse on the study of natural philosophy* (1831) the English polymath John Herschel (1792–1871) drew the distinction between observation and experiment as follows:

> We have thus pointed out to us, as the great, and indeed only ultimate source of our knowledge of nature and its laws, EXPERIENCE... But experience may be acquired in two ways: either, first, by noticing facts as they occur, without any attempt to influence the frequency of their occurrence, or to vary the circumstances under which they occur; this is OBSERVATION: or, secondly, by putting in action causes and agents over which we have control, and purposely varying their combinations, and noticing what effects take place; this is EXPERIMENT. (Herschel 1831, 76)

John Stuart Mill made the same distinction in his equally influential *System of Logic* (Mill [1843] 1974, 382), and so did William Stanley Jevons (1835–1882) in his *Principles of Science* from 1874.

> We are said to *experiment* when we bring substances together under various conditions of temperature, pressure, electric disturbance, chemical action, &c., and then record the changes observed. Our object in inductive investigation is to ascertain exactly the group of circumstances or conditions which being present, a certain other group of phenomena will follow. If we denote by A the antecedent group, and by X subsequent phenomena, our object will usually be to discover a law of the form A = AX, the meaning of which is that where A is X will happen. (Jevons 1920, 416)

For a modern formulation of the distinction we can turn to Jürgen Habermas:

> In an experiment we bring about, by means of a controlled succession of events, a relation between at least two empirical variables. This relation satisfies two conditions. It can be expressed *grammatically* in the form of a conditional prediction that can be deduced from a general lawlike hypothesis with the aid of initial conditions; at the same time it can be exhibited *factually* in the form of an instrumental action that manipulates the initial conditions such that the success of the operation can be controlled by means of the occurrence of the effect. (Habermas 1968, 162; 1978, 126)

The older, wider meaning of "experiment" that includes passive observation has not disappeared. Well into the 1930s many writers of academic textbooks used the term "experimental" as synonymous with "empirical" (Winston and Blais 1996).[3] And outside of academia the distinction between a scientific experiment and a scientific observation is often befogged. In 1993 the Israeli Ministry of Transportation raised the interurban speed limit from 90 to 100 kph and in 1995 further to 110 kph. This was done in order to reduce transportation time, adjust speed limits to observed actual speeds, and harmonize Israeli regulations with those of the European Union. The raise in speed limits was announced by the director of the Road Safety Authority as an "experiment", but it was not carried out as a scientific experiment with study design, protocol, principles for evaluation, and ethical safeguards such as cut off points for early termination. Concerned physicians had demanded that the "experiment" be subjected to the same type of ethical review as medical experiments on humans, but they were not listened to. Not surprisingly, the resulting higher velocities led to an increased death toll on Israeli roads (Richter et al. 2001; Hansson 2011).

In the remainder of this essay I will need the distinction that Herschel, Mill, Jevons, Habermas and others have expressed with the two words "observation" and "experiment". I will follow precedent and use these well-established terms to make the distinction. In other words, by an experiment I will mean a procedure in which some object of study is subjected to interventions (manipulations) that aim at obtaining a predictable outcome or at least predictable aspects of the outcome. Predictability of the outcome, usually expressed as repeatability of the experiment, is an essential component of the definition. Experiments provide us with information about regularities, and without predictability or repeatability we do not have evidence of anything regular. A procedure that we carry through can only have the intended function of an experiment in our deliberations if its setup is so constructed that it determines at least some aspects of the outcome.[4]

Obviously, what is important here is the distinction, not the term we use to express it. In contexts where it is infeasible or undesired to reserve the term "experiment" for the concept I use it for here, the phrase "controlled experiment" can be used instead for that concept.[5]

Two Types of Experiments

In addition to the distinction between experiments and non-experimental observations we need to distinguish between two types of experiments, namely directly

[3] Even today, promoters of so-called experimental philosophy use the term "experiment" about questionnaires and other studies that behavioural scientists would classify as observational non-experimental studies. For a criticism, see Hansson (2014).

[4] Obviously, it need not be known beforehand which aspects of the outcome are determined by the setup, or in particular how they are determined by it.

[5] The term "scientific experiment" is not useful for the purpose since it excludes controlled experiments performed in a non-scientific setting, for instance in the traditions among farmers and craftspeople referred to in Sect. 3.

action-guiding experiments and epistemic experiments.[6] An experiment is directly action-guiding if and only if it satisfies the following two criteria:

(1) The outcome looked for should consist in the attainment of some desired goal of human action, and
(2) the interventions studied should be potential candidates for being performed in a non-experimental setting in order to achieve that goal.

These criteria are satisfied for instance in a clinical trial. In a clinical trial of an analgesic the outcome looked for is efficient pain reduction with minimal negative side effects. Therefore, the experimental intervention is a treatment that might be administered to achieve this outcome in patients in an ordinary clinical setting. Other examples of directly action-guiding experiments are agricultural field trials, many technological tests such as tests of the longevity of light bulbs, and social experiments trying out the effects of different methods of social work. In contrast, an epistemic experiment aims at providing us with information about the workings of the world we live in. Therefore, the outcome looked for is one that provides such information, and it need not coincide with anything that a sensible person would wish to happen except as part of the experiment itself.

Both historical and philosophical accounts of experiments and experimental method have been almost exclusively devoted to epistemic experiments in science, and surprisingly little has been written on directly action-guiding experiments. Beginning with the historical issue—that I will only treat very briefly here[7]—the exclusion of directly action-guiding experiments from scholarly attention has consigned most of the early history of experiments to oblivion. Contrary to what is commonly assumed in historical studies, the origin of experimentation was neither academic nor curiosity-driven. Instead, it was unacademic and driven by practical needs.

Agriculture, the crucial technological innovation that made civilization possible, was the result of thousands of years of extensive and continuous experimentation (Bray 2000). Records from all parts of the world confirm that indigenous and traditional farmers do indeed experiment (Earls 1998; Chandler 1991; Johnson 1972). The Mende people in Sierra Leone have a special word, hungoo, for such experiments. A hungoo can for instance consist in planting two seeds in adjacent rows, and then comparing the output in order to determine which seed was best (Richards 1989).

Just like farmers, craftspeople of different trades have performed directly action-guiding, i.e. technological, experiments already in prescientific times. As one example of this, extensive experiments on the composition of glass were performed in the early Islamic period in Raqqa (Ar-Raqqah) in eastern Syria. Analysis of

[6] The distinction was introduced in Hansson (2015). Strictly speaking, the distinction is not between different types of experiments but between different types of interpretations of experiments, viz. the interpretation of experiments for action-guiding or epistemic purposes. However, since most experiments are purposeful only for one of the two types of interpretation the convenient locution of two types of experiments will be used here.

[7] See Hansson (2015) for a more extensive treatment.

artefacts and debris from the eighth to eleventh centuries has revealed clear signs of experimentation, including a so-called chemical dilution line in order to search systematically for optimal proportions of the main ingredients (Henderson et al. 2004, 2005). There are strong reasons to believe that systematic experimentation to optimize batch recipes must also have taken place in many other places and at various times throughout the world, for instance to optimize alloys, mortars, and dyes (Malina 1983; Moropoulou et al. 2005). We also have strong indications that the increasingly slim structures of the cathedrals from the High Gothic period (around 1140–1350) are the outcomes of advanced experiments. The pillars and other construction elements of a new cathedral were made somewhat slimmer than those of its predecessors. Various signs such as cracks in the mortar showed the builder where there was a need for flying buttresses or other auxiliary supporting structures (Mark 1972, 1978; Wolfe and Mark 1974).

All this adds to a context against which we can understand the experimental developments in Renaissance science. Thanks to the devoted scholarship of Zilsel (1941, 1942, 2000) we know that skilled craftsmen had a major role in the development of experimental science in the early Renaissance. Their technical skill was obviously needed to make the experimental equipment; in addition they had experiences from a long tradition of experimental methodology that had been developed for directly action-guiding purposes but could now be used in epistemic experiments as well.

Justifying Experiments

Why should we perform experiments? Why manipulate nature when we can instead study its undisturbed behaviour? Or as J.E. Tiles puts it: "How can we possibly learn the principles which govern the action of natural bodies if we do not let nature take its course?" (Tiles 1993) Today this is not much discussed but in the early modern period it was an important issue in natural philosophy. Thomas Hobbes was one of the more prominent opponents of the experimental, i.e. manipulative approach to the study of nature (Shapin 1985, 1996, 110–111).

But let us pose this somewhat unfashionable question. When should we make experiments, and when should we instead observe our objects of study without manipulating them? I would like to propose that the answer to that question depends on whether the investigation has an action-guiding or an epistemic purpose.

Action-Guiding Experiments

In the cases when the purpose of the investigation is to directly guide action, I want to defend a rather strong claim: *For action-guiding purposes, an experiment is always epistemically preferable*[8] *to a non-experimental observation.* The reason for

[8] It may not be preferable tout court, for instance if there are ethical reasons not to perform the experiment.

this is that directly action-guiding experiments have a strong and immediate justification:

> *The immediate justification of directly action-guiding experiments*
> If you want to find out whether you can achieve Y by doing X, do X and find out whether Y occurs.

This is so self-evident that philosophical justification appears difficult. In particular, we have no problem accounting for the usefulness of performing an intervention or manipulation (namely X). We want to know the effects of such an intervention, and then it is much better to actually perform it than for instance to passively observe the workings of nature without performing the intervention. If you wish to find out which of two electric water boilers is the fastest, fill them both with the same amount of water of the same temperature, start them at the same time and wait to see which of them finishes first. Similarly, if you want to know whether a decreased intake of salt reduces blood pressure, perform an experiment with volunteers willing to eat less salt in order to see the actual effects of such a change in diet.

Someone might ask: Is this really true? Can directly action-guiding experiments really be that strongly justified? Have we not learned that all experiments are theory-laden? The answer is that indeed we have, but what we have not been told is that the theory-ladenness of an experiment comes with its epistemic interpretation and is not relevant for a directly action-guiding interpretation. Therefore, directly action-guiding experiments are remarkably theory-independent.[9] You may have all kinds of reasons for believing that a drug helps against a particular disease, for instance biochemical reasons, belief in someone's authority, or belief in the supernatural powers of the plant that the drug was made from. But if a competently performed experiment shows that the drug does not have the effect in question, then that is a result you cannot reasonably dismiss.

Not surprisingly, the strongly supported, theory-independent information from action-guiding experiments is unwelcome among proponents of theories that may have to be revised or deserted due to that information. Ideological rejections of the outcomes of directly action-guiding experiments are particularly prominent in the medical area. When clinical trials, i.e. directly action-guiding treatment experiments, were first introduced they had great difficulties in gaining acceptance. Before the twentieth century only few such experiments were performed, and their outcomes were often resisted by the medical authorities. It was only well into the twentieth century that therapeutic experiments became the established method to determine the best treatment method for diseases. A decisive step forward was taken in 1948 with the first publication of a clinical trial using modern methodology, including the randomization of patients between treatment groups (Marshall et al. 1948; Doll 1998; Cooper 2011). But the introduction of clinical trials was not without resistance. The opposition has been strongest in psychiatry (Williams and

[9] They are of course not completely theory-independent. My claim is (only) that they are radically less so than epistemic experiments, and in fact not more theory-dependent than any non-empty statement about empirical subject-matter.—It should also be noted that theory-ladenness refers to the interpretation of experiments rather than to their physical execution. Therefore, strictly speaking, the distinction made here concerns action-guiding versus epistemic interpretations of experiments. Cf. footnote 6.

Garner 2002), which is unsurprising due to the strong influence in that specialty of theory-based doctrines prescribing what treatments to use. But in recent years, experimental methodology has taken big steps forward in psychiatry. The remaining pockets of resistance in the healthcare sector can now be found among so-called alternative therapists who commonly reject directly action-guiding experiments, claiming incorrectly that the effects of their treatments are invisible in clinical trials. (For a critical appraisal of such claims, see Jerkert 2013.) A philosophical account of experimentation that separates out action-guiding experiments and clarifies their strong justificatory status can contribute to better use of experimental studies in social contexts such as healthcare and public health where reliable action-guiding information is urgently needed.

Epistemic Experiments

For investigations with an epistemic purpose, experiments do not have the strong prerogative that they have in the directly action-guiding case. There are many scientific issues for which non-experimental observation is the preferred method. If as a biologist you want to know how the grey heron feeds its offspring, it would not be a good idea to expose it to different experimental conditions. Instead you would have to find a way to observe its spontaneous behaviour without disturbing it. Similarly, if as a social scientist you want to investigate the decision-making of the Supreme Court, then you have no reason to expose its members to experimental conditions, since what you want to know is what they actually do in the conditions they are actually working under, not what they would do under various other conditions. Generally speaking, if you want to know how nature or humans behave spontaneously, then non-experimental observation rather than experiment is the preferred type of study. Many branches of science are based on well-established forms of non-experimental observations, such as anatomical dissections, botanical inventories, geological surveys, and demographic statistics.

If, on the other hand, you are looking for knowledge about the connections between different types of events, then experiments are typically the preferred type of study. In other words, it is in these cases that it can be useful to manipulate your study object. If you want to know how the parental feeding behaviour of the grey heron is influenced by the availability of food, it will be useful to perform an experiment in which you artificially increase its access to food. (Do this in the wild, not in a laboratory!) If you want to know how the decisions of judges are influenced by various characteristics of the defendant, you can perform an experiment in which judges are asked to make judgments in various hypothetical cases.

With this restriction to cases in which we are searching for connections rather than spontaneous behaviour, why should we (in these cases) perform experiments rather than non-experimental observations?

An important partial answer (or rather condition for any other answer that we may want to give) is that the same regularities ("laws") that govern the spontaneous workings of nature also apply when nature responds to human intervention. This is something that most of us take for granted today, but for previous generations it was not always self-evident. Even Mill saw reasons to state it explicitly:

> For the purpose of varying the circumstances, we may have recourse (according to a distinction commonly made) either to observation or to experiment; we may either *find* an instance in nature suited to our purposes, or, by an artificial arrangement of circumstances, *make* one. The value of the instance depends on what it is in itself, not on the mode in which it is obtained... (Mill [1843] 1974, 381)

This assumption puts experimentation on a par with non-experimental observation, but that is not quite enough for a justification. What reasons are there why experiments could be superior to other observations? Mill provides us with one such reason:

> Thus far the advantage of experimentation over simple observation is universally recognised: all are aware that it enables us to obtain innumerable combinations of circumstances which are not to be found in nature, and so add to nature's experiments a multitude of experiments of our own. (Mill [1843] 1974, 382)

Essentially the same argument was put forward already by Francis Bacon, in somewhat more drastic terms:

> For like as a man's disposition is never well known till he be crossed, nor Proteus ever changed shapes till he was straitened and held fast; so the passages and variations of nature cannot appear so fully in the liberty of nature as in the trials and vexations of art. (Bacon ([1605] 1869, 90; cf. Zagorin 1998, 61–62)

Even in Bacon's formulation this is a rather diffident justification of experimentation. It does not defend the specific characteristic of experiments, namely the active element, the intervention that distinguishes it from mere observations. It only defends experimentation in terms of an advantage that experiments often but not always have over other observations, namely the advantage of providing data from a much larger number of combinations of circumstances.

But Mill also adds a second justification of experimentation:

> When we can produce a phenomenon artificially, we can take it, as it were, home with us, and observe it in the midst of circumstances with which in all other respects we are accurately acquainted. If we desire to know what are the effects of the cause A, and are able to produce A by means at our disposal, we can generally determine at our own discretion, so far as is compatible with the nature of the phenomenon A, the whole of the circumstances which shall be present along with it: and thus, knowing exactly the simultaneous state of everything else which is within the reach of A's influence, we have only to observe what alteration is made in that state by the presence of A. (Mill [1843] 1974, 382)

In modern terms: In an experiment we can control all other factors than the one which we wish to change in order to determine its effects on the outcome.

To this I would like to add a third justification of experimentation, namely its special relation to our conceptualization of causality. Let us first note the distinction between two meanings of causality. First, it can refer to cause–effect relationships. These are binary production relationships such that if *C* is a cause of the effect *E*, then, in the absence of contravening circumstances, if *C* takes place then so does *E*. Secondly, by causality we can mean the totality of regularities in the universe, or its workings. These two senses are often taken to coincide. To know how something works and to know the cause–effect relationships that determine its operations would seem to be one and the same thing. That, however, is an oversimplification. The universe does not seem to work according to binary cause–effect relationships. This was seen clearly by Bertrand Russell, who noted that the cause–effect pattern does not capture Newtonian physics, in which movements emerge from complex interactions among a large number of bodies, all of which influence each other simultaneously (Russell 1913, 1). For instance, the planetary system cannot be accounted for in terms of chains of successive cause–effect relationships, but it can be adequately described and predicted in a model where the total effects of a large number of simultaneous interactions are obtained as the solution of a system of differential equations. Today, this also applies to central problems in many other areas of science, such as climatology, economics, and biological population dynamics. In these and many other areas, an account restricted to binary cause–effect relationships will lack much of the explanatory power of modern science. But although the cause–effect model of the universe is problematic, we have a strongly entrenched tendency to employ it. When trying to understand the world we want to do so in terms of cause–effect relationships. This is where experiments come in. Experiments are constructed to show how a certain intervention produces a certain outcome. The relationship between intervention and outcome is fairly close to a cause–effect relationship. This cognitive fit contributes to explaining why experiments tend to be so useful in our strivings to understand the world.

Concluding this, we have three reasons to employ epistemic experiments to find out the general connections among events in the world: We can study a larger number of combinations of circumstances than under unmanipulated conditions, we can achieve better control over these circumstances, and we can couch our investigations in the terms of cause–effect relationships which is how we are disposed to think about the world.

The main conclusions of this section are summarized in Fig. 1 that shows how the optimal mode of empirical investigation depends on what we want to know.

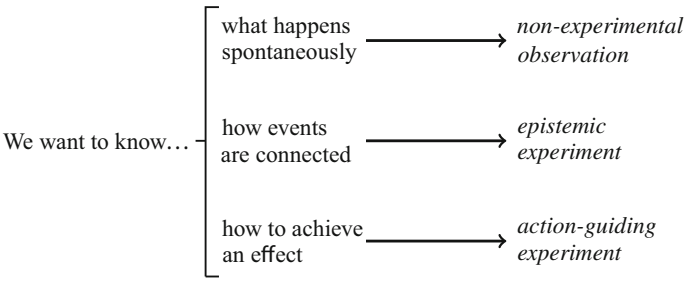

Fig. 1 The choice of observational methods in search of different types of knowledge

The Best is Not Perfect

Obviously, choosing the best method does not guarantee a reliable result. Both observation and experimenting are difficult activities, full of pitfalls. Mistakes are common in the execution and interpretation of all types of empirical investigations, both experimental and non-experimental (observational) ones. Some of these mistakes are specific for a particular type of investigation, for instance failures in the particular type of instrumentation. Here I will focus on the types of imperfections that are common to most if not all empirical investigations. These imperfections originate either in the behaviour of the observanda (objects of observation) or in the behaviour of the investigators.

Beginning with the observanda, even in non-experimental observations we usually assume some regularity in that which we observe. Suppose that after observing of a couple of grey herons for several days I conclude that the female bird does most of the feeding of the nestlings. This may be a too rash conclusion; perhaps this male is untypically lazy or a bad fisher, and other males do a much larger share of the feeding. This is an example of *variability* of the observanda. It is equally important in experimental investigations (of both the directly action-guiding and the epistemic variety). It would for instance be unwise to draw any conclusions from a drug test on only a single patient since there may be variations in how the body responds to the drug, and this patient may be untypical.

In addition to being irregular, observations may also be regular but in a way that confuses us. Suppose that someone compares the grades of schoolchildren with their clothes, and finds that children with cheap clothes have worse grades. It would in all probability be a mistake to interpret this correlation as proof of a direct causal connection between clothes and grades. There is of course an underlying factor, namely family income, that can explain the difference. Factors that create such spurious regularities are called *confounders*.

Turning to the investigators, *evaluator bias* is a problem in all types of observational studies, both experimental and non-experimental ones. Usually, it takes the form of "seeing what we expect". If a physician believes a new drug to be efficient, then she tends to pay more attention to improvements than deteriorations in the health of patients who take the drug. If a botanist believes that a particular species grows predominantly close to water she may take more note of its occurrence in such places, etc.

These three sources of error, variability, confounders, and evaluator bias, are common to non-experimental and experimental studies. In experiments, the experimenter has a double role; she is not only observer of the effects produced but also implementor of the experimental conditions that give rise to these effects. Therefore, a fourth source of error makes its appearance in experiments, namely *implementor bias*. Consider a clinical trial in which a physician distributes different drugs to patients that have been assigned to different experimental treatments. If the physician believes one of the drugs to be better than the others, then this may influence her behaviour towards the patient in various ways, both in terms of physical and psychological treatment. Since we are presumably looking for the effects of the drugs, not those of the physician's beliefs about the drugs, this is a

misleading effect that we want to eliminate from our interpretation of the experiment.

These four sources of error all reduce the reliability of the experiment or observation, i.e. the degree to which we are justified in relying on its outcome.[10] Whether it provides us with knowledge we can use also depends on our previous knowledge and on the questions we ask. The major problem in this respect for directly action-guiding experiments is *unrealism*, i.e. that either the intervention or the conditions under which it is performed in the experiment differ from the situations for which action guidance is sought. For instance, clinical trials performed on carefully selected and diagnosed patients in a specialist clinic do not always tell us everything we need to know about the effects of administering the same treatment in general practice. For epistemic experiments, the major problem is instead *unconnectedness* with existing theories and with new theoretical issues that can be opened up for investigation. This applies equally to non-experimental investigations with epistemic purposes.

The six shortcomings in (experimental and non-experimental) observations are summarized in Fig. 2.

Improvement Strategies

The practical art of performing empirical investigations in science consists to a large part in coping with the four major sources of error described in the previous section. Coping can of course consist either in eliminating the effect or separating it out so that it does not influence the interpretation of the observation. In what follows I will describe seven major coping strategies and track their historical origins as far as possible.

Multiple Observations

Compare the following two eleventh century proposals for treatment experiments:

> In order to evaluate the efficacy of ginseng, find two people and let one eat ginseng and run, the other run without ginseng. The one that did not eat ginseng will develop shortness of breath earlier. (Claridge and Fabian 2005, 548)

> [It] is like our judgement that the scammony plant is a purgative for bile; for since this [phenomenon] is repeated many times, one abandons that it is among the things which occur by chance, so the mind judged that it belongs to the character of scammony to purge bile and [the mind] gave into it, that is, purging bile is an intrinsic characteristic belonging to scammony. (McGinnis 2003, 317)

[10] The reliability of an experiment is closely related to its repeatability.

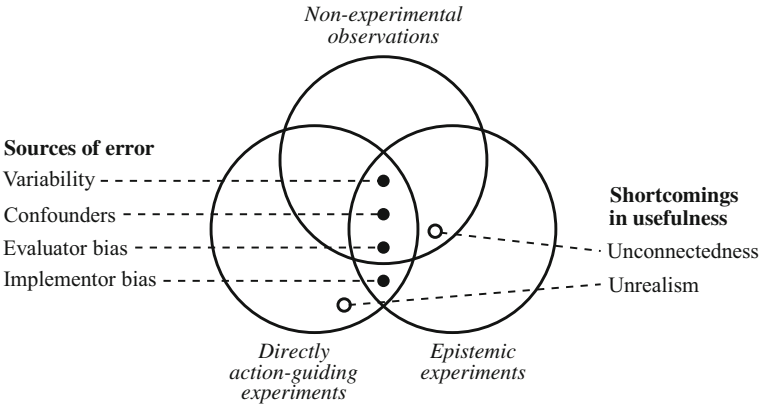

Fig. 2 The major types of shortcomings in experimental and non-experimental observations

Both statements are by great polymaths of the period, the first by the Chinese Su Song (1020–1101) and the second by the Persian Avicenna (Abd Allah ibn Sina, c. 980–1037). It is not known if any of these two proposals was actually carried out, but at any rate Avicenna's experiment is superior since it requires repetition "many times". If the person selected in Su Song's experiment for taking ginseng actually became breathless later than the other runner then there is no way to know whether this was due to ginseng or to the natural variation in short-windedness among human beings. The example shows that multiple observations can be used as a means to counteract the uncertainty created by variability in the observanda. This applies to non-experimental as well as experimental observations.

The use of multiple observations is one of the experimental strategies that was available in pre-scientific, directly action-guiding experimentation. For instance, farmers performing agricultural experiments do not sow a single seed but enough of them to cover a small plot. The reason for this is of course the variabilities among seeds in germination and outcome that are well-known to farmers. Similarly, the amount of debris from glass-making experiments in Raqqa indicates that repetition was part of the experimental strategy.

Comparisons

Although Su Song, as quoted in the previous subsection, proposed too few experimental subjects (only two) in his ginseng trial, he did at least realize that he had to make a comparison. In order to find out whether ginseng prevents shortness of breath it is not sufficient to just observe some person(s) who have taken the drug; you must compare to person(s) who have not taken it. This elementary insight applies, of course, to non-experimental observations as well.

The need for comparisons gives us reason to expand the recipe for directly action-guiding experiments that was given above, namely:

> If you want to find out whether you can achieve Y by doing X, do X and find out whether Y occurs.

In some cases this recipe can be used since we have the background knowledge that without X, Y will not occur. But if such knowledge is not available, or can be put to doubt, then we need to arrange for an experimental comparison:

> If you want to find out whether you can achieve Y by doing X, both do and refrain from doing X under similar circumstances, and find out whether the cases with X differ from those without X in that Y occurs more often or to a higher degree.

Such a comparison is called a control experiment or a control arm of the experiment, and the group of objects in the control experiment is called a control group. Several of our examples above show that the use of controls goes back to pre-scientific, directly action-guiding experiments. Early agricultural experimenters compared alternative seeds or treatments, and glass-makers compared alternative batch mixes. Today control groups are a self-evident part of the experimental design in all areas of science. This applies both to directly action-guiding experiments (such as a clinical trial) and epistemic experiments (for instance when the effect of light on a chemical reaction is tested by performing it both in the dark and under exposure to light).

The comparative method has been further developed into the method of *parameter variation*. By this is meant that the experiment is performed repeatedly with systematic variation of one or several variables. For instance, if we wish to determine the effects of temperature on a chemical reaction, we perform the experiment at different temperatures. Alhazen (965-c.1040) has been credited with performing the first experiment with parameter variation. (In an optical experiment he varied the size of an aperture through which moonlight was projected.) (Schramm 1963, 287) However, we have already noted that experiments with systematic variation of the proportions of major ingredients in glass were performed in Raqqa in the early Islamic period, and that in all probability, systematic changes in batch composition had been performed long before that. We can conclude that parameter variation, just like the control experiment, has its origin in pre-scientific, directly action-guiding experimentation.

Measurement

One of the most important safeguards against evaluator bias is a precise methodology for describing the outcome. This applies to both non-experimental and experimental observations.

A farmer who wants to compare the yield of a new cultivar to that of the variety he currently grows will probably not be satisfied by the general impression he gets from harvesting plots with the two cultivars. Instead he will collect the harvests from equally sized plots with the two varieties and gauge them with some measurement vessel. Measurement is an efficient way to avoid evaluator bias. Today it is used almost universally in epistemic experiments—most scientific

laboratories have a plethora of devices that are employed to assign numerical values to various parameters. (There is also another reason for this. In addition to being an antidote to evaluator bias, measurement also provides the means necessary to determine numerical relationships that can be compared to mathematical models of the workings of nature.)

Measurement, in particular of weight and length, was well established for technical and other purposes long before modern science. It is also plausible that such devices were used in technological experiments. However, I have not been able to find historical evidence of outcome measurement in pre-scientific, action-guiding experiments. What we do know, however, is that the methods of measurement used in scientific experimentation were largely developed from measurement techniques that had been developed for technological purposes.

Blinding

Another way to avoid evaluator bias is to divest the evaluator of the information that triggers her bias. Consider a radiologist who is going to evaluate the radiological development of a disease in two groups of patients that have undergone different treatments. If she knows which patients have received which treatment, then her expectations of the efficacy of the two treatments may influence her judgement. This can be avoided by simply withdrawing that information from her. This is called *blinding* (or masking). It is used extensively in clinical trials and often also in other research areas, both in experimental and non-experimental observations.

In addition, blinding can be used in experimental settings against the other type of experimenter bias, namely implementor bias. For instance, the patients in a clinical drug trial should all meet a physician. If the physician knows which patients receive which drug, then that may influence her interactions with the patients in various ways, and these interactions may have impact on the patient's health status. The standard way to avoid this is to keep the physician ignorant about which patients receive which drug.

The history of observational and experimental blinding remains to be written. It has been claimed that Dom Pérignon (1639–1715), a French monk and wine maker, performed blind-testing of wines in order not to be influenced by the expectation effects on his judgment (Bullock et al. 1998), but I have not been able to verify that claim in reliable historical sources. In 1817, Stradivarius violins were compared to other violins in blind hearing tests (Quatremre de Quincy 1817; Fétis 1868, 249). The earliest scientific study with blinded evaluators seems to have been that performed in 1784 by a commission of the French Academy of Sciences led by Benjamin Franklin that had been assigned to investigate Franz Mesmer's claims of animal magnetism. Under blinded conditions, mesmerists were unable to distinguish which objects had gone through an occult procedure described as filling them with vital fluid. The commission concluded from these blinded experiments that the fluid had no physical existence (Sutton 1981; Lopez 1993). Blinding was taken up by critics of deviant belief systems and used rather extensively in the nineteenth century in investigations of extraordinary claims. One interesting example is the experiments performed in the 1830s by Michel Eugène Chevreul (1786–1889) in

which he showed that the effects of dowsing disappeared under blinded conditions. Another such example is a double blind (and randomized) test of homeopathy performed in Nuremberg in the same decade, showing—equally unsurprisingly—that a homeopathic drug had no other effects than pure water (Stolberg 2006).

But for a long time, blinding was almost only used when there was suspicion that an effect was due to suggestibility or fraud. After World War II awareness of the fallability of human judgment became widespread among researchers, not least due to influence from psychological research. As a consequence of that, blinding has become an increasingly common means to avoid the dangers of evaluator and implementor bias (Kaptchuk 1998). Judging by the evidence I have been able to find, blinding is a fairly late addition to the arsenal of safeguards in empirical research, and it seems to have its origin in modern science rather than in pre-scientific experimentation.

Randomization

One of the forms that implementor bias can take is selection bias. Suppose that you are going to test a drug by comparing two groups of patients, one of which receives the drug and the other not. The results of the test are not of much value if you (unconsciously) assign patients in better health to one of the two groups, and those in worse health to the other. The preferred method to avoid this error is randomization, i.e. letting chance decide which patients are assigned to which group.

As far as I have been able to glean from the available literature, randomization is, just like blinding, a rather recent addition to the methodological arsenal. It seems to have its origin in modern science rather than in pre-scientific experimentation. In its modern form it was developed for experimental agriculture by the statistician Ronald Fisher (1890–1962) in the 1920s when working at the Rothamsted Experimental Station in England. He developed statistical methods for agricultural trials that included the random assignment of cultivars to fields. The method began to be used in clinical medicine in the late 1940s. In 1948 the first report employing the method was published (Marshall et al. 1948; Doll 1998). Since then it has spread to a large number of other research areas, including psychology, the social sciences, and experimental biology (where animals rather than humans are distributed randomly between treatment groups).

Statistical Tests

Finally, let us return to the virtually ubiquitous problem of variability. Common sense suffices to tell us that the effects of ginseng on shortness of breath cannot be determined by comparing just one person who takes the drug and one who does not. The method of multiple observations can be used to reduce (or rather even out) the effects of variability. But how many observations are necessary to obtain reliable results? Perhaps quite a few. The Flemish physician and scientist Jan Baptist van Helmont (1580–1644) came, intuitively, close to a modern understanding of this in a challenge concerning the efficacy of different medical treatments:

> Let us take out of the hospitals, out of the Camps, or from elsewhere, 200, or 500 poor People, that have Fevers, Pleurisies, &c. Let us divide them into halfes, let us cast lots, that one half of them may fall to my share, and the other to yours; . . . we shall see how many funerals both of us shall have: *But* let the reward of the contention or wager, be 300 florens, deposited on both sides. (quoted from Armitage 1982)

What we want to avoid is to mistakenly treat an effect as real that is in fact due to random variation. Obviously, the risk of being "fooled" by chance effects cannot be completely eliminated, but we can make it improbable. In order to do so we have to refrain from drawing conclusion from outcomes that would easily arise by chance alone. But unfortunately, our intuitions about what this requires for instance in terms of sample size are far from reliable (Sedlmeier and Gigerenzer 1997). Statistical tests developed in the last 100 years or so have made it possible to deal with this problem in a systematic way. For this final item on my list of safeguards, the historical evidence is quite clear and in no need of being treated at length: These tools were not available in prescientific experimental traditions. They have been developed within science.

Summary

The six safeguards are summarized in Fig. 3, where they are also tabulated against the four general pitfalls in empirical research that they are used to counteract. All these safeguards are needed in both directly action-guiding and epistemic experiments, and most of them also in non-experimental observations. Three of them, namely multiple observations, comparisons, and measurements, appear to have been part of the methodologies and ways of thinking that scientific experimentation took over from pre-scientific, directly action-guiding experimentation. The other three, namely blinding, randomization, and statistical tests, have their origins in modern science.

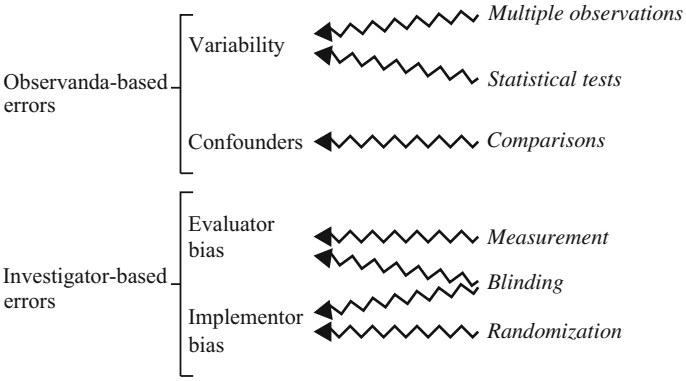

Fig. 3 The four major general pitfalls in experimental inquiry and six major safeguards to cope with them

Conclusion

I hope to have substantiated the importance of distinguishing experiments, in the standard scientific sense of the term, from non-experimental observations. Both are needed, but for different purposes. Scientific experimentation builds on traditions from pre-scientific technological experiments, and important parts of the methodological arsenal of modern experimental science can be traced back to that tradition.

In conclusion, experimentation involves a unique combination of acting and observing. This is a combination with a unique knowledge-creating capacity that is used every day in scientific laboratories but whose epistemological foundations are in need of further investigations. In particular, the justification of epistemic (i.e., not directly action-guiding) experiments depends on assumptions on the relationships between manipulation, causality, and the regularities of nature that are in need of additional, detailed investigations.

References

Armitage, P. (1982). The role of randomisation in clinical trials. *Statistics in Medicine*, *1*, 345–352.

Bacon, F. ([1605] 1869). *The Advancement of Learning* [*Of the Proficience and Advancement of Learning, Divine and Human*]. Oxford: Clarendon Press.

Bray, W. (2000). Ancient food for thought. *Nature*, *408*(9), 145–146.

Bullock, J. D., Wang, J. P., & Bullock, G. H. (1998). Was Dom Perignon really blind? *Survey of Ophthalmology*, *42*, 481–486.

Chandler, P. M. (1991). The indigenous knowledge of ecological processes among peasants in the People's Republic of China. *Agriculture and Human Values*, *8*, 59–66.

Claridge, J. A., & Fabian, T. C. (2005). History and development of evidence-based medicine. *World Journal of Surgery*, *29*, 547–553.

Cooper, M. (2011). Trial by accident: Tort law, industrial risks and the history of medical experiment. *Journal of Cultural Economy*, *4*(1), 81–96.

de Quincy, A. C. Q. (1817). Institut de France, *Le Moniteur universel*, August 22 1817, no 234, p. 924.

Doll, R. (1998). Controlled trials: The 1948 watershed. *BMJ. British Medical Journal*, *317*(7167), 1217–1220.

Earls, J. (1998). The character of Inca and Andean Agriculture. Accessed 29 July 2013. http://macareo.pucp.edu.pe/jearls/documentosPDF/theCharacter.PDF

Eastwood, B. S. (1968). Mediaeval empiricism: The case of Grosseteste's optics. *Speculum: A Journal of Mediaeval Studies*, *43*, 306–321.

Fétis, F.-J. (1868). *Biographie Universelle des Musiciens et Bibliographie Générale de la Musique. Tome 1* (2nd ed.). Paris: Firmin Didot Frères, Fils, et Cie.

Habermas, J. (1968). *Erkenntnis und interesse*. Frankfurt am Main: Suhrkampf Verlag.

Habermas, J. (1978). *Knowledge and human interests* (J. J. Shapiro, Trans). London: Heinemann. (2nd edition).

Hansson, S. O. (2011). Do we need a special ethics for research? *Science and Engineering Ethics, 17*:21–29.

Hansson, S. O. (2013). Defining pseudoscience – and science. In M. Pigliucci & M. Boundry (Eds.), *The Philosophy of Pseudoscience* (pp. 61–77). Chicago: Chicago University Press.

Hansson, S. O. (2014). Beyond experimental philosophy. *Theoria, 80*:1–3.

Hansson, S. O. (2015). Experiments before science? - what science learned from technological experiments. In S. O. Hansson (Ed.), *The role of technology in science. Philosophical perspectives*. Dordrecht: Springer.

Henderson, J., McLoughlin, S. D., & McPhail, D. S. (2004). Radical changes in Islamic glass technology: Evidence for conservatism and experimentation with new glass recipes from early and middle Islamic Raqqa, Syria. *Archaeometry*, *46*(3), 439–468.

Henderson, J., Challis, K., OHara, S., McLoughlin, S., Gardner, A., & Priestnall, G. (2005). Experiment and innovation: Early Islamic industry at al-Raqqa, Syria. *Antiquity*, *79*(303), 130–145.

Herschel, Wi. (1831). *A preliminary discourse on the study of natural philosophy*, part of Dionysius Lardner, *Cabinet Cyclopaedia*. Chicago.

Jerkert, J. (2013). Why alternative medicine can be scientifically evaluated: Countering the evasions of pseudoscience. In M. Pigliucci & M. Boudry (Eds.), *The philosophy of pseudoscience* (pp. 305–320). Chicago: Chicago University Press.

Jevons, W. S. (1920). *The principles of science: A treatise on logic and scientific method*. London.

Johnson, A. W. (1972). Individuality and experimentation in traditional agriculture. *Human Ecology*, *1*(2), 149–159.

Kaptchuk, T. J. (1998). Intentional ignorance: A history of blind assessment and placebo controls in medicine. *Bulletin of the History of Medicine*, *72*(3), 389–433.

Klein, U. (1996). Experiment, Spiritus und okkulte Qualitäten in der Philosophie Francis Bacons. *Philosophia Naturalis*, *33*(2), 289–315.

Klein, U. (2005). Experiments at the intersection of experimental history, technological inquiry, and conceptually driven analysis: A case study from early nineteenth-century France. *Perspectives on Science*, *13*, 1–48.

Lopez, C.-A. (1993). Franklin and Mesmer: An encounter. *Yale Journal of Biology and Medicine*, *66*(4), 325–331.

McEvoy, J. (1982). *The philosophy of Robert Grosseteste*. Oxford: Clarendon Press.

McGinnis, J. (2003). Scientific methodologies in Medieval Islam. *Journal of the History of Philosophy*, *41*, 307–327.

Malina, J. (1983). Archaeology and experiment. *Norwegian Archaeological Review*, *16*(2), 69–78.

Mark, R. (1972). The structural analysis of Gothic cathedrals. *Scientific American*, *227*(5), 90–99.

Mark, R. (1978). Structural experimentation in Gothic architecture: Large-scale experimentation brought Gothic cathedrals to a level of technical elegance unsurpassed until the last century. *American Scientist*, *66*(5), 542–550.

Marshall, G., et al. (1948). Streptomycin treatment of pulmonary tuberculosis: A medical research council investigation. *British Medical Journal*, *2*(4582), 769–782.

Mill, J. S. ([1843] 1974). *A system of logic ratiocinative and inductive*. In J. M. Robson (Ed.), *Collected works of John Stuart Mill*, Vol. VII. Toronto: University of Toronto Press.

Moropoulou, A., Bakolas, A., & Anagnostopoulou, S. (2005). Composite materials in ancient structures. *Cement & Concrete Composites*, *27*, 295–300.

Pesic, P. (1999). Wrestling with Proteus: Francis Bacon and the 'torture' of nature. *Isis*, *90*, 81–94.

Richards, P. (1989). Farmers also experiment: A neglected intellectual resource in African science. *Discovery and Innovation*, *1*, 19–25.

Richter, E. D., Barach, P., Berman, T., Ben-David, G., & Weinberger, Z. (2001). Extending the boundaries of the declaration of Helsinki: A case study of an unethical experiment in a non-medical setting. *Journal of Medical Ethics*, *27*, 126–129.

Robison, W. (2008). Hume and the experimental method of reasoning. *Southwest Philosophy Review*, *10*(1), 29–37.

Russell, B. (1913). On the notion of a cause. *Proceedings of the Aristotelian Society*, *13*, 1–26.

Ryle, G. (1949). *The concept of mind*. Chicago: The University of Chicago Press.

Ryle, G. (1971 [1946]). Knowing how and knowing that. In: G. Ryle, *Collected papers*, volume 2. (pp. 212–225). New York: Barnes and Nobles.

Schramm, M. (1963). *Ibn al-Haythams Weg zur Physik*. Wiesbaden: Franz Steiner Verlag.

Sedlmeier, P., & Gigerenzer, G. (1997). Intuitions about sample size: The empirical law of large numbers. *Journal of Behavioral Decision Making*, *10*, 33–51.

Shapin, S. (1985). *Leviathan and the air-pump: Hobbes, Boyle, and the experimental life*. Princeton, NJ: Princeton University Press.

Shapin, S. (1996). *The scientific revolution*. Chicago: University of Chicago Press.

Stolberg, M. (2006). Inventing the randomized double-blind trial: The Nuremberg salt test of 1835. *Journal of the Royal Society of Medicine*, *99*, 642–643.

Sutton, G. (1981). Electric medicine and mesmerism. *Isis*, *72*(3), 375–392.

Tiles, J. E. (1993). Experiment as intervention. *British Journal for the Philosophy of Science*, *44*(3), 463–475.

Wiesemann, C. (1991). *Josef Dietl und der therapeutische Nihilismus: zum historischen und politischen Hintergrund einer medizinischen These*, volume 28 of *Marburger Schriften zur Medizingeschichte*. Frankfurt am Main: Peter Lang.

Williams, D. D. R., & Garner, J. (2002). The case against 'the evidence': A different perspective on evidence-based medicine. *British Journal of Psychiatry*, *180*, 8–12.

Wolfe, M., & Mark, R. (1974). Gothic cathedral buttressing: The experiment at Bourges and its influence. *Journal of the Society of Architectural Historians, 33*(1):17–26.

Zagorin, P. (1998). *Francis Bacon*. Princeton: Princeton University Press.

Zilsel, E. (1941). The origin of William Gilberts scientific method. *Journal of the History of Ideas*, *2*, 1–32.

Zilsel, E. (1942). The sociological roots of science. *American Journal of Sociology*, *47*, 544–562.

Zilsel, E. (2000). The social origins of modern science. In D. Raven, W. Krohn, & R. S. Cohen (Eds.), *Boston studies in the philosophy of science* (Vol. 200). Dordrecht: Kluwer Academic Publishers.