DENNY BORSBOOM, GIDEON J. MELLENBERGH and JAAP VAN HEERDEN

FUNCTIONAL THOUGHT EXPERIMENTS

ABSTRACT. The literature on thought experiments has been mainly concerned with thought experiments that are directed at a theory, be it in a constructive or a destructive manner. This has led some philosophers to argue that all thought experiments can be formulated as arguments. The aim of this paper is to draw attention to a type of thought experiment that is not directed at a theory, but fulfills a specific function within a theory. Such thought experiments are referred to as functional thought experiments, and they are routinely used in applied statistics. An example is given from frequentist statistics, where a thought experiment is required to establish the probability space. It is concluded that (a) not all thought experiments can be formulated as arguments, and (b) the role of thought experiments is more general and more important to scientific reasoning than has previously been recognized.

1. INTRODUCTION

It could be argued that all science begins with counterfactual thinking. For the most basic question of inquiry, 'why is the world as it is?', can only originate from the idea that the world could have been different. That is, an explanation need only be considered if there are phenomena to be explained, and phenomena require explanation only if they are not taken for granted. Not taking phenomena for granted requires one to consider the possibility that the world could have been different, which is only possible upon the consideration of counterfactual alternatives. This, in effect, means one has to perform a thought experiment – where a thought experiment is loosely defined as a line of reasoning that proceeds from counterfactual premises.

In the light of the importance of counterfactual reasoning and thought experimenting to such basic issues of inquiry, it seems somewhat surprising that the role of the thought experiment in science has for long been neglected by philosophers of science. Apart from the pioneering work of Mach (1905/1976), and a paper by Kuhn (1977), it has only been for the last decade that a considerable body of conceptual research has emerged on the thought experiment as a philosophical, mathematical, and scientific strategy (Brown 1991; Horowitz and Massey 1991; Sorensen 1992; Wilkes 1988).

Synthese **130:** 379–387, 2002.

One of the most clarifying achievements in this emerging body of literature is the taxonomy proposed by Brown (1991). He classifies thought experiments as being *destructive* (aimed at refutation of a theory), *constructive* (providing support for a theory), or *platonic* (destructive for all theories but one). An example of a destructive thought experiment is Einstein's refutation of Maxwell's theory of light. Einstein reasoned that, if Maxwell's theory were correct, he would have to see a light beam as a spatially oscillatory electromagnetic field at rest, when running at the speed of light. The thought experiment is destructive, because it is used to derive a contradiction from the premises of Maxwell's theory. An example of a constructive thought experiment is the silicon-brain experiment, in which it is argued that, if all the neurons in your brain were gradually replaced by computerchips, you would still be conscious after the replacement. This thought experiment has been used as an argument for functionalism by various philosophers (see, for example, Dennett 1991). Finally, an example of a platonic thought experiment is Galileo's famous refutation of Aristotle's theory of motion. Aristotle's theory stated that heavier objects fall with greater acceleration than lighter objects. Galileo reasoned that, if Aristotle's theory were true, a heavier object tied to a lighter object should, when falling from a given height, simultaneously reach the ground sooner and later than the heavier object alone. This thought experiment is platonic in the sense that it does not only refute Aristotle's theory, but at the same time establishes a single alternative theory, namely the theory that acceleration does not depend on the mass of an object. Platonic thought experiments may therefore be conceived of as destructive thought experiments that are constructive for a single alternative. The majority of thought experiments that have thus far been considered in the literature can be classified in Brown's taxonomy, a possible exception being what Bunzl (1996) has called the consistency thought experiment. The essential feature of consistency thought experiments is that 'typically, such thought experiments result in a modification of background assumptions rather than any change in the theory itself' (p. 234). An example is the Einstein-Podolsky-Rosen (EPR) thought experiment, which eventually did not serve to refute or support quantum mechanics, but resulted in a modification of our background assumptions.

All thought experiments that have thus far been considered in the literature are of the type that is directed at a theory, be it in a constructive or a destructive manner (notwithstanding the fact that consistency thought experiments do not result in a modification of the theory, the EPR thought experiment was clearly directed at quantum theory). This has led some philosophers, such as Norton (1991), to claim that all thought experiments

can be formulated as arguments. The aim of the present paper is to draw attention to a type of thought experiment that is not directed at theories, and therefore cannot be formulated as an argument, nor subsumed under Brown's taxonomy. These thought experiments may best be characterized as functional, in the sense that they create a conceptual framework that allows for the application of a theory. A functional thought experiment that we will consider in some depth is employed routinely in the application of frequentist statistics in order to establish a probability space. As such, it plays an important role in frequentist statistics as well as in the areas where statistics is applied. Before we discuss the characteristics of the thought experiment, we will shortly sketch the frequentist conception of probability, and illustrate the problem by an example drawn from the theory of mental testing.

2. THE FREQUENTIST CONCEPTION OF PROBABILITY

It is not extremely difficult to provide a syntax for probability theory in the form of a calculus. Axiomatized systems have, for example, been provided by Kolmogorov (1933) and Rényi (1970). However, the interpretation of the probability syntax is not straightforward, and to some extent arbitrary. Several interpretations have survived up to this day (see Nagel 1936, or Fine 1973, for an overview of possible interpretations). The most influential interpretation of probability in applied statistics is the so-called frequentist account, which is in terms of long run frequencies.

The frequentist's long run interpretation of probability is seemingly uncomplicated. Toss a coin infinitely many times, and the limiting value of the relative frequency of heads in these trials is the probability of heads. For many applications it is important that these trials satisfy an independence condition. It is, however, not easy to specify what 'independent' means before probability itself has been defined. We cannot use the concept of probability to define the independence of trials, because probability itself is defined in terms of these trials. Hacking (1965) escapes the vicious circle by deducing the independence of trials from the independence of the outcomes of these trials. Following Hacking's brand of frequentism, one requires that the trials are 'unrelated' and defines the probability of an outcome as the relative frequency of that outcome. Then the concept of probability can be applied to define the statistical independence of trials, from which the independence of trials themselves can be deduced.

The definition of probability as the limit of a relative frequency in an infinitely long run of independent observations yields a very general framework for the application of the probability calculus. Its general applic-

ability, together with its 'objective' character, have made the frequentist conception the most widely held view in applied statistics.

3. THE IMAGINARY LONG RUN: LORD AND NOVICK'S BRAINWASH

The frequentist idea of probability as relative frequency in the long run is, upon closer examination, problematic. It is certainly intuitively plausible for games of dice, but it is not at all straightforward for many areas in which statistics is needed, and indeed has proven successful. We will illustrate this statement by an analysis of the problem as it occurs in mental testing, because it clarifies what the problem with the long run actually is.

In the theory of psychological testing, a basic assumption is that observed test scores contain measurement error. A subject's score on a psychological test will be influenced by factors that are not of interest to the researcher. Some of these are systematic (to be understood as stable over time, for example, certain personality characteristics), and some are random (i.e., accidental, for example, the subject had a headache at the testing occasion). Here, we will be concerned only with the random part of the error.

The basic idea of classical test theory (Lord and Novick 1968) is that there exists a 'true score', which is to be conceived of as the observed score, stripped of its random error. The true score T is thus defined as the observed score O minus the random error E. This leads to the classic equation $O = T + E$. Of course, this definition is empty unless some procedure is specified to define what error actually is. This procedure is borrowed from the theory of random errors as developed in astronomy (Edgeworth 1888; see also Stigler 1986, and Hacking 1990). The idea is that, if we take measurements on many occasions, it is plausible to define the true score as the expectation of the observed scores over repeated measurements, so that $T = \mathcal{E}(O)$. This is unproblematic in the context of astronomical measurements, where the repeated observations can reasonably be assumed to be independent (in the sense of the previous section). Consequently, the frequentist conception of probability as long-run relative frequency can be utilized. It is then reasonable to define the true score as the expectation over repeated observations, which is, by its definition, a constant.

This line of reasoning fails in psychological testing. Disregarding the fact that performing many measurements on the same subject is unrealistic, people learn, get tired, become familiar with the testing procedure, and so on. As a consequence, trials are not unrelated and the outcomes of the trials will not, in general, be independent. So, Hacking's (1965) method of

deducing the independence of trials from the independence of outcomes does not work here. However, if we want to apply a long run frequency interpretation of probability, trials must be independent.

Lord and Novick solve this problem by introducing a thought experiment originally proposed by Lazarsfeld (1959).

> Suppose we ask an individual, Mr. Brown, repeatedly whether he is in favour of the United Nations; suppose further that after each question we 'wash his brains' and ask him the same question again. Because Mr. Brown is not certain as to how he feels about the United Nations, he will sometimes give a favorable and sometimes an unfavorable answer. Having gone through this procedure many times, we then compute the proportion of times Mr. Brown was in favor of the United Nations. (Lord and Novick 1968, pp. 29–30)

In the thought experiment, the observations are rendered independent as a result of the brainwashing procedure. Now we may apply the frequency interpretation of probability. It then becomes possible to take the expectation of the observed scores and to define this expectation as the true score, which is again a constant. In the particular case of Mr. Brown, the expectation equals the probability of him giving a favorable answer, which is estimated by the proportion of times he was in favor of the United Nations. The introduction of this thought experiment has proven extremely useful in the development of both classical (Lord and Novick 1968) and modern (Hambleton and Swaminathan 1985) test theory.

Lord and Novick are very plain in admitting that they use a thought experiment. They are forced to do so by their subject matter. A long sequence of repeated testing occasions is so obviously implausible that they cannot ignore the problem. The thought experiment, however, is not symptomatic for psychological testing. It is implicitly present in many applications of frequentist statistics. A long run of independent observations on the same unit does not exist anywhere in the real world. Almost independent, yes; practically independent, yes; truly independent, no. The notion of independent observations is an idealization, although it often is a useful assumption (it would certainly be a pathological case of hair-splitting to criticize the assumption of independent trials in throwing dice). In virtually every application of inferential statistics, however, the thought experiment is needed. Hacking (1965; p. 10) hinted at this when he said that long run frequency is concerned with 'what the long run frequency is or would be or would have been'. In order to be able to invoke the expectation of the outcome of measurements in medicine, economics, or social research, one always has to talk about 'what the long run frequency would have been if . . . '. The conditional part of the sentence contains, in these cases, counterfactual premises. It is a thought experiment.

4. THE NATURE OF STATISTICAL THOUGHT EXPERIMENTS

The frequentist thought experiment is different from those used in physics or philosophy. The primary chracteristic that distinguishes the thought experiment from the type that has received attention in the literature so far, is that it is not directed at any theory in particular. Although the thought experiment is necessary for employing the frequentist scheme of statistical inference, it is not used to support the frequentist view (in the sense of showing that the frequentist theory is 'true'). Rather, it is an integral part of that view: It creates the conceptual framework rather than supporting it. This type of thought experiment may be best characterized as functional: A functional thought experiment is not aimed at refuting or supporting a theory, but has a specific function within a theory. In the case of frequentist statistics, it functions as a semantic bridge, providing a real world interpretation for the abstract syntax of probability.

The distinction between functional and constructive/destructive thought experiments runs parallel to the distinction between theory and model. A theory can be true or false: a constructive or destructive thought experiment is intended to show that it is true or false. A statistical model, if it is internally consistent, can only be shown to be applicable or not applicable; and the functional thought experiment is used to convince the reader that it actually is applicable. This, in the frequentist thought experiment, is established by an appeal to analogy. The frequentist conception of probability is generally deemed applicable to games of dice, and as a result the outcome of trials in such a game can be considered a random variable, satisfying the required independence conditions. In the Lord and Novick thought experiment, we are asked to imagine a situation where a subject's response could be considered analogous to the outcome of trials in throwing dice. The crucial point is that an important structural characteristic (the 'randomness' of the trial outcomes) is preserved in the new domain. If this much is granted, the rest of the theory follows smoothly.

The use of functional thought experiments is not limited to the semantic bridge function in frequentist statistics. Actually, there are many statistical models and techniques that are not interpretable without a thought experiment. Because it is not within the scope of this paper to give a complete and thorough overview of the use of functional thought experiments, we only mention briefly some of the models in which they are used. One example is the causal model of Rubin (1974; see also Holland 1986), in which it is necessary to consider a concept called the counterfactual expectation. In a standard experimental setup employing an experimental and a control condition, this would for example be the counterfactual expectation of the

dependent variable in the control condition, that would have 'existed' if the subjects in that condition had been assigned to the experimental condition. A related statistical technique where thought experiments are needed is the use of covariates as control variables. In this technique, the observed means for a given variable are corrected for a covariate. For example, a mean difference between men and women on annual income may disappear, once the observed difference is corrected for the sex difference in educational level. Such a result is interpreted as 'there would not have been a sex difference in annual income, had men and women had the same average eductional level', which is clearly a counterfactual statement. All these statistical thought experiments are functional, since they render a model applicable but are not directed at a theory.

5. DISCUSSION

Functional thought experiments are not aimed at a theory, but create a conceptual framework for the application of a theory or model. Therefore, they cannot be formulated as arguments in the sense of Norton (1991). Another consequence is that functional thought experiments cannot be described as constructive or destructive for a theory, and are not captured in the taxonomy of Brown (1991). It therefore seems that the functional thought experiment specifies a distinct class of thought experiments. This may be one of the reasons that it has gone unnoticed in the philosophical literature on thought experiments. Another reason may be that statistical thought experiments are, in most cases, not explicitly presented as counterfacual lines of reasoning. Most treatments of statistics do not explicate the counterfactuals that are employed in statistical arguments, Lord and Novick (1968) being one of the rare exceptions to this rule.

There are several implications of the present discussion that are of interest to the role of thought experiments in science. In the literature on thought experiments, it is generally contended that the method of thought experiment is used almost exclusively in philosophy and physics. Upon the present discussion, however, this does not seem to be the case. Thought experiments are used incidentally in physics and regularly in philosophy, but they are commonplace in medicine, biology, and the social and behavioral sciences. Confidence intervals, p-values, reliabilities, and likelihood ratios all result from the same kind of thought experiment: The counterfactual long run. We therefore think it is not unreasonable to say that the long run frequency thought experiment, although generally not recognized as such by those who employ it, is the most common thought experiment in science. The role of thought experiments and counterfactual reasoning

may therefore be more general and more important to the development of science than has been previously recognized.

ACKNOWLEDGEMENT

The authors would like to thank Susanne Hammen for some useful discussions on thought experiments in psychometrics.

REFERENCES

Brown, J. R.: 1991, *The Laboratory of the Mind: Thought Experiments in the Natural Sciences*, Routledge, London.

Bunzl, M.: 1996, 'The Logic of Thought Experiments', *Synthese* **106**, 227–240.

Dennett, D. C.: 1991, *Consciousness Explained*, Little Brown, Boston.

Edgeworth, F. Y.: 1888, 'The Statistics of Examinations', *Journal of the Royal Statistical Society* **51**, 598–635.

Fine, T. L.: 1973, *Theories of Probability*, Academic Press, New York.

Hacking, I.: 1965, *Logic of Statistical Inference*, Cambridge University Press, Cambridge.

Hacking, I.: 1990, *The Taming of Chance*, Cambridge University press, Cambridge.

Hambleton, R. K. and H. Swaminathan: 1985, *Item Response Theory: Principles and applications*, Kluwer-Nijhoff, Boston.

Holland, P. W.: 1986, 'Statistics and Causal Inference', *Journal of the American Statistical Association* **81**, 945–959.

Horowitz, T. and G. J. Massey (eds): 1991, *Thought Experiments in Science and Philosophy*, Rowman and Littlefield, Savage, MD.

Kolmogorov, A.: 1933, *Grundbegriffe der Warscheinlichkeitsrechnung*, Springer, Berlin.

Kuhn, T. S.: 1977, 'A Function for Thought Experiments', in T. S. Kuhn, *The Essential Tension*, The University of Chicago Press, Chicago, IL.

Lazarsfeld, P. F.: 1959, *Latent Structure Analysis*, McGraw-Hill, New York.

Lord, F. M. and M. R. Novick: 1968, *Statistical Theories of Mental Test Scores*, Addison-Wesley, Reading, MA.

Mach, E.: 1905/1976, *Knowledge and Error*, Reidel, Dordrecht.

Nagel, E.: 1936, *Principles of the Theory of Probability*, The University of Chicago Press, Chicago, IL.

Norton, J.: 1991, 'Thought Experiments in Einstein's Work', in T. Horowitz and G. J. Massey (eds), *Thought Experiments in Science and Philosophy*, Rowman and Littlefield: Savage, MD.

Rényi , A.: 1970, *Foundations of Probability*, Holdaen-Day, San Francisco.

Rubin, D.: 1974, 'Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies', *Journal of Educational Psychology* **66**, 688–701.

Sorensen, R.: 1992, *Thought Experiments*, Oxford University Press, Oxford.

Stigler, S. M.: 1986, *The History of Statistics*, Harvard University Press, Cambridge, MA.

Wilkes, K.: 1988, *Real People*, Clarendon Press, Oxford.

Department of Psychology
Faculty of Social and Behavioral Sciences
University of Amsterdam
Roetersstraat 15
1018 WB Amsterdam
The Netherlands
E-mail: ml_borsboom.d@macmail.psy.uva.nl