RMM Vol. 4, 2013, 1–19
http://www.rmm-journal.de/

Volker Gadenne

External Validity and the New Inductivism in Experimental Economics*

Abstract:
The idea of external validity, which is well-known in the social sciences, has recently also been emphasized in experimental economics. It has been argued that external validity is an important criterion in experimental research, which has been neglected by philosophy of science. In connection with this criterion, a methodology has been advanced in which inductive generalization and analogical inference play a central role. The hypothetico-deductive methodology is said to be untenable, or at least insufficient. In this paper, hypothetico-deductivism is defended. The idea of external validity, and the new plea for inductivism, is critically discussed. It is pointed out that the fundamental problems of inductivism are still unsolved. The criterion of external validity is superfluous and misleading. And the problems in experimental research associated with external validity can well be solved on the basis of deductivism.

1. Introduction

When a new scientific discipline evolves, its members often become interested in methodological questions, and some of them address such questions in articles and textbooks. In the last decade, this can be found in experimental economics where problems of validity, theory testing, deductive versus inductive thinking etc. were discussed. It is argued that the traditional methodology of economics does not sufficiently account for the goals and problems of the new experimental discipline and has to be complemented by new rules. Two textbooks of this kind, which have become widely recognized, are *The Methodology of Experimental Economics* by Francesco Guala (2005) and *Experimental Economics* by Nicholas Bardsley et al. (2010).

In this discussion of methodological questions, some experimental economists have emphasized a conception of validity that is seldom mentioned in philosophy of science but well-known in psychology and the social sciences: the theory of *internal* and *external validity*. Introduced by Campbell (1957) and Campbell und Stanley (1963), this conception has become a central part of the methodology of the social sciences (Gadenne 1976).

* For helpful comments, I am grateful to Max Albert and Jakob Kapeller.

Guala (2005) refers to Campbell und Stanley. Especially, he stresses external validity, and criticizes general philosophy of science for having neglected this important criterion. He discusses external validity in connection with the controversy between deductivism and inductivism. Guala declares Popper's deductivism as untenable, and pleads for inductive thinking, which gives external validity a central role in experimental research:

> "The growth of scientific knowledge, then, seems to proceed by careful induction from the particular to the particular, rather than by means of bold theoretical generalizations. Theories do play a role in the design of experiments and at various stages in the process of inductive inference, but at both the starting and the end points of the inference, we find knowledge of rather specific systems." (Guala 2005, 201f.)

Bardsley et al. put forward a similar view though they are less dismissive of the hypothetico-deductive method. They also emphasize external validity but allow for hypothetico-deductive and inductive approaches to develop in partnership:

> "[. . .] we are beginning to see the emergence of an alternative—or perhaps complementary—methodology for economics, in which inductive generalization plays a central part." (Bardsley et al. 2010, 194)

In the following, I want to analyze and criticize the idea of external validity and the new embracement of inductive thinking. I will argue that this development results from a too narrow view of the hypothetico-deductive method, and a rather vague understanding of induction. There is no need to introduce external validity and inductive rules. Induction is untenable in all its versions, and external validity is an equivocal, misleading idea. The problems of experimental research in economics can well be accounted for by deductivism, if the latter is properly understood.

The views I am going to criticize can be found in the literature cited above, and in many other textbooks. However, I want to emphasize from the beginning that I consider the books of Guala and Bardsley et al. as well worth reading. They contain a lot of valuable discussions of economic experiments. But the authors' proposals concerning induction and external validity are unconvincing.

2. Two Kinds of Validity in Empirical Research

In an experiment, a variable X, usually called the treatment or the independent variable, is manipulated. It is observed whether another variable Y, the dependent variable, changes its value. Extraneous variables, that is, other factors which could also affect Y, are controlled, e.g., by elimination, or by keeping them constant. We might be interested, e.g., whether attitude change is affected by a

communicator's credibility. Here Y is the subjects' attitude toward some object. The subjects are given some relevant information by a communicator (X) who appears to them either credible (condition 1) or untrustworthy (condition 2).

Assume the result of such an experiment is a covariation between X and Y. There was significantly more attitude change when the communicator was perceived as credible. Then, according to the theory of internal and external validity, two questions arise. The first is whether the change in X is the *cause* of the change in Y. "Did in fact the experimental treatments make a difference in this experimental instance?" (Campbell and Stanley 1963, 175) This is called the question of *internal validity*. It is 'internal' insofar as it is restricted to the subjects and situation of the experiment that was performed.

Now empirical research usually aims at the discovery of general regularities and laws. If a singular causal relation is detected in an experiment, we want to know whether the same holds outside that experiment. This is the problem of *external validity*. "External validity asks the question of generalizability. To what populations, settings, treatment variables, and measurement variables can this effect be generalized?" (Campbell and Stanley 1963, 175)

What exactly is it that is valid or invalid? It is the conclusion drawn from the data of a given experiment. Internal validity means that it is justified to conclude from the data of a given experiment that there is a causal relation between X and Y in this experiment. External validity means that it is justified to generalize the causal relation from the subjects in this experiment to a certain population, and to other settings and variables.

It is also common to call experiments or experimental designs themselves valid or invalid. But this usually means that the conclusions drawn from the results are valid or invalid.

I have explained internal and external validity with respect to experiments. However, these criteria can be applied to nonexperimental studies as well, like quasi-experiments or field studies. Generally speaking, they can be applied to any empirical research that aims at the discovery or test of causal relations and general regularities. They are not relevant if one is only interested in describing a singular situation.

It is often said that there is a trade-off between internal and external validity. In order to enhance internal validity, one has to arrange a situation in which everything is under control. The laboratory is suitable for that purpose. However, such an artificial environment is much unlike the situations in which people act in normal life. Hence laboratory experiments are judged as internally valid but externally less valid, and are often criticized for that. By contrast, field studies are said to be externally valid, but to lack internal validity since it is difficult in natural settings to control extraneous variables.

3. Internal Validity, External Validity, and the Problem of Induction

Consider the following experimental design. On a group of persons, observations of variable Y (e.g., blood pressure) are recorded at two times, t_1 and t_2. Between t_1 and t_2, the subjects receive a treatment (e.g., a new medicament). This design is regarded as internally invalid, since it does not enable the control of other influences that may occur between t_1 and t_2.

By contrast, the following design is internally valid. Subjects are randomly distributed in two groups. In the experimental group they receive a medicament, in the control group a placebo. Afterwards Y is measured in both groups. If there is a significant difference in Y between the two groups, it is justified to conclude that X had a causal effect.

Internal validity requires control of extraneous variables. That means one should use the best control techniques available for the given purpose. Control is never perfect. Therefore, internal validity gives no guarantee that the inferred causal statement is true. Internal validity means that it is reasonable or rationally justified to accept a causal statement though one cannot be absolutely sure it is true.

There is a remarkable agreement on the question which designs are sufficient for internal validity, and which are not. With respect to *external validity*, the situation is quite different. To what populations, settings, treatment variables, and measurement variables may we generalize the causal relation found in this experiment? Assume we have found that the communicator regarded by the subjects as credible was more effective in causing immediate attitude change than the communicator regarded as untrustworthy. The subjects were 50 male students of psychology. Does this causal effect also hold for female students, for students of other disciplines, for less educated people, or for any further individual at all? Furthermore, does it hold for other communicators, other attitudes, and other methods of measuring attitudes? With respect to experiments in economics we might ask whether the result of a certain study also holds for people who do not know the theories of economics, or for subjects who are given the opportunity to gain more experience with the tasks used in the experiment.

We may well ask these questions. But nobody can answer them just with the help of the data from the experiment whose external validity is to be assessed. Campbell and Stanley (1963, 175) conceded that "the question of external validity, like the question of inductive inference, is never completely answerable [. . .]". However, there seems to be no 'incomplete' answer either.

Methodologists who regard external validity as important usually refer to *representativeness*. In order to justify a generalization from a sample to a population, the sample should be representative of that population. This is also stressed by Guala (2005, 197): "This is, in fact, the logic underlying the best-known external validity control—representative sampling. If you want to generalize to population B, you should make sure that you have in the lab good

representatives of the individuals in B (you need students if you want to generalize to students, housewives for housewives, mammals for mammals, etc.)."

However, this does not solve the problem. Of course we should study mammals if we want to make general statements about mammals. Representativeness in this sense can and must be realized. But in order to justify a generalization, we would need representativeness in another sense: Every feature relevant for the causal relation found in the experiment should be distributed in the sample exactly in the same way it is distributed in the population. We do not know whether this is the case. Random sampling does not provide a solution. It is possible to draw a random sample from a small finite population, but not from an open class, such as the class of all mammals, or all humans. Experiments are usually performed to answer questions about laws, and laws are statements about open classes of individuals.

The problem of external validity is the problem of *induction*: Can it be justified to conclude that a universal statement is true, or probably true, by assuming that certain observation statements are true? Popper (1959, chapter I) argued, with reference to Hume, that it is not possible to justify such inferences from observations to laws or theories. An inductive inference cannot be founded on deductive logic. And the attempt to justify induction by arguing that induction has proven successful in the past is circular. An inference from past success to future success is itself an inductive inference.

Recently, some philosophers have advocated Bayesianism (Howson and Urbach 1989), which can be interpreted as a special form of induction. However, Bayesianism is itself faced with serious problems (Albert 2003). Furthermore, it cannot provide a solution to the problem of external validity. Bayesianism requires that, previous to empirical evidence, probabilities are assigned to hypotheses. As we shall see in a moment, the problem of external validity does not even arise if one starts with hypotheses.

Thus answering the question of external validity amounts to solving the problem of induction. It should be noted, however, that not all generalizations are made in order to prove some universal hypothesis. If we have repeatedly observed that A was followed by B, we may generate a universal hypothesis without claiming that this hypothesis is proven to be true or probable by the observations that gave rise to its formulation. In this case no problem of induction arises. Generalizing as a means to generate hypotheses raises no difficulty.

However, someone who asks for external validity is not just interested in generating a hypothesis. The concept of validity clearly implies the idea of justifying something, and the proposition which is to be justified here is a universal proposition, or at least a proposition about other persons and situations than those observed so far.

4. Theory Testing without External Validity

The problem of generalization and external validity arises if one makes observations which are not yet related to a universal hypothesis or theory. Now assume we start with a universal hypothesis or theory. A theory T is tentatively proposed in order to solve a problem, e.g., to explain some facts. From T, together with some additional assumptions A (singular statements, auxiliary hypotheses), a test prediction P is derived. Observations are made. If the observational results agree with P, we say that T has been *confirmed*. But we do not inductively conclude that T has been proven true. If the observations deviate from P, and we accept non-P, we have to conclude that T & A is false. This poses the question which proposition from T & A should be given up, a question known as Duhem's problem. I shall discuss this problem below. Now I first want to point out that in theory testing no question of generalizability comes up.

If the test result is non-P, there is nothing to be generalized. The same holds if the result is P. Especially, there is no question *to what population the result should be generalized*. If we regard T as confirmed by P, there is no open question concerning the 'population' of our study. The population T refers to has been fixed from the beginning. In sciences searching for regularities and laws, a theory T consists of some universal statements. A universal statement refers to an open class of objects, e.g., the class of all physical objects, all living beings, or all humans. This class is called the *domain* of T. For a theory to be testable, it must be clear what its domain is. Thus, the only 'population' involved here is the domain of T.

By contrast, the problem of internal validity appears as relevant also from the perspective of theory testing. Tests of causal hypotheses require that the effects of different causes are kept apart from each other, and this has to be done with the help of control techniques. Let H be the hypothesis that A always leads to B. We realize an instance of A in an experiment and predict B. If B occurs we regard H as confirmed by this result. However, this conclusion is justified (internally valid) only if we took care that B was not caused by an event other than A. If B was caused by C, we might wrongly conclude that H is confirmed even if H is false. The experiment must be arranged in such a way that other possible causes of B are controlled.

Now if there is, from the perspective of theory testing, no 'external' problem to be solved, it makes no sense to speak of an 'internal' problem either. Internal validity simply becomes the problem of control of extraneous variables.

Let us return to external validity and generalization. If generalization was the only way to achieve general knowledge, that is, knowledge of regularities and laws, and if generalization is rejected as a method of justification, this would lead to skepticism in regard to general knowledge. The method of theory testing solves this problem. We can gain general knowledge by testing theories. Such knowledge consists of theories that have been confirmed.

But does it really matter whether the theory or the empirical evidence comes first? Let E be the result from some empirical study, and H a hypothesis that

fits E. Is it really important whether H is formulated prior to E or after E? And if E is taken to confirm H in the former case, doesn't E confirm H equally well if H was constructed by generalizing E?

These questions are controversially discussed in philosophy of science. Here is the main argument for the view that observational results which are known when the hypothesis is formulated cannot confirm this hypothesis. Confirmation, properly understood, is not a purely formal relation. It depends, partly, on pragmatic aspects. There is a fundamental difference between the situations 'H prior to E' and 'E prior to H'. If E is known first, H is constructed in such a way that it fits E. If, by contrast, H is formulated before one knows the result of the study, the result could be non-E, which would count as evidence against H. So H is put at risk, it is seriously tested. The risk is particularly severe if E is highly unexpected in the light of the available knowledge. If E occurs nevertheless, this confirms H. Confirmation is here conceived of as an unexpected positive outcome. On the other hand, if we knew E in advance and construct H by generalizing E, the result non-E is excluded from the beginning. Nothing could happen in such a study that contradicts H. H is not put at risk, and should therefore not regarded as confirmed by E.

The methodology I have cited here is known as the *hypothetico-deductive* methodology. In the following, I shall shortly call it *deductivism*. There are different versions of deductivism. The best-known is Popper's, which is also called *falsificationism*. Falsificationism has been extensively discussed in philosophy of science, and has been declared as untenable by its critics. Since I have presented a form of deductivism as a solution to the problem of general knowledge, and as a means to avoid the difficulties created by inductive thinking, I have to address two further questions. Is deductivism a tenable methodology? And is deductivism sufficient as a methodology of economics and the social sciences, or should it be complemented by inductive rules (as the new inductivists claim)?

5. In Defense of Hypothetico-deductivism

In his *Logik der Forschung* from 1934, Popper declared as the criterion of empirical science the *falsifiability* of theories. Scientific theories should be formulated in such a way that observations can contradict them. Moreover, they should, in every conceivable way, be exposed to tests which could refute them. Today many philosophers of science believe that Popper's methodology is untenable. It is argued that falsificationism cannot solve Duhem's problem. It is also objected that falsificationism is incompatible with the way famous scientists like Galilei and Newton proceeded. Two of Popper's main critics who argued from the history of science were Thomas Kuhn (1962) and Paul Feyerabend (1975).

I cannot here present in detail Popper's early view (usually called *falsificationism*), his later view (*critical rationalism*), the various objections by its critics, and the answers given by Popper and other critical rationalists. Instead I want to present and defend an improved version of the hypothetico-deductive

methodology. It is part of Popper's later view, and is held by many other critical rationalists (e.g., Hans Albert 1968; Musgrave 1999). I concentrate on the point that was central in the controversy over deductivism, the questions associated with the idea of *falsification.*

Assume a theory T is to be tested. From T & A, a test prediction P is derived. P should be not just any logical consequence from T & A. It should be a consequence that is not yet known to be true. If we knew in advance that P is true, we would not carry out a genuine test. We would only pretend to test T. Deductivism insists that tests have to be genuine in this sense, and recommends that, if possible, they should be *severe*. A test is *severe* if P is predicted by T but expected to be false in the light of the available background knowledge, or some rival theory. Severe tests are highly informative since they help to find, and sort out, false assumptions. In most cases, they lead to the result that some assumption is false and has to be corrected (Gadenne 2002). If the result is non-P, we learn that the new theory may contain an error we have to identify. If P results, we learn that something we have held to be true (background assumptions, or an established theory T'), must be questioned.

If T is tested and P is accepted as true, T is said to be confirmed. This does not mean that T has been proven as true with certainty. It does not mean either that T is true with high probability or with a probability that can be exactly specified. (Popper used the term 'corroboration' to distinguish his concept from Carnap's concept of 'inductive confirmation', which was conceived as a probability.) Yet we assume that theories can be more or less confirmed. The degree of confirmation depends of the number and kinds of tests T has passed. A severe test of T contributes to T's degree of confirmation more than a normal (genuine) test. And if T has been confirmed by an empirical study, repeating that study does not give T much additional confirmation, whereas a study of a new kind does (provided the result is P). Theories should therefore be tested in different kinds of situations.

Does confirmation justify accepting a theory as true? There is a small group of critical rationalists who vehemently deny this (Miller 1994). However, most scientists and philosophers, including those who hold deductivism, think that confirmation makes a theory more plausible and more acceptable. I here also adopt the idea that confirmation increases the acceptability of a theory. And if T has passed severe tests, and if there is no rival theory that is also confirmed, it is justified to accept T tentatively as true (Musgrave 1999, chapter 16).

The last point has to be emphasized because some critics reject critical rationalism for not allowing any positive evaluation of theories. For example, Guala (2005) rejects Popper's methodology for that reason. But deductivism is usually (and here) understood as a view that includes principles of falsification as well as principles of acceptance.

Now assume that T is tested and non-P is accepted as true. This is a *falsification* of T & A. Most importantly, this does not mean that we can be sure that T is false. Nor does it mean that we necessarily have to reject T. We can instead decide to question P, or propose that an assumption from A is false. In principle,

all statements are to be regarded as *fallible*, as possibly false. This holds for theories as well as for observation statements. Therefore, a falsification can never be final or conclusive. The judgment that T has been falsified only means that non-P has been tentatively accepted, so that T & A must be (tentatively) judged as false. And if, in addition, A is (tentatively) accepted, T must be (tentatively) judged as false.

When an empirical result is taken to falsify a theory, this result is not usually a description of one single observation. In most cases, it is a statement that summarizes many observations. Often it is itself a low-level hypothesis that covers the observations made so far. For example, if T is the Copernican theory, the empirical result that falsifies T could be the statement that the planet Mars describes an orbit around the sun that deviates from an exact circle.

Deductivism allows us to defend a theory against contradicting results. This must be allowed since auxiliary hypotheses and even observations might be false and sometimes have indeed turned out to be false in the history of science (Chalmers 1999, 87–91). Admittedly, this carries the danger that false theories might be defended endlessly by their advocates, which would prevent progress. Therefore, deductivism proposes that anyone who defends a theory T must point out which other assumption from A should be questioned and tested. One could argue, e.g., that certain measuring instruments did not function correctly. If this can be confirmed, the burden is no longer on T. But if no disfunctioning is detected, T is still to be regarded as (tentatively) falsified. Especially, it is not allowed to save theories by putting the blame on unknown and untestable disturbances.

But how can we confirm or falsify anything, and make progress, if any statement is conceived as fallible? This is a decisive question. Deductivism rests on the epistemological assumption (a weak form of empiricism) that people are more often false in theorizing than in perceiving. Our hypotheses and theories are often false. Perceptions may also be false. But if we arrange a situation in which the object to study can be very well observed, our observation statements are rather reliable. Scientists argue about theories and explanations, but they seldom disagree that, e.g., a piece of blue litmus paper has turned red, or that someone has marked the answer 'yes' in the first item of a questionnaire. This is why observations are used to test theories, and not vice versa.

How can Duhem's problem be solved? Duhem demonstrated that in science a test prediction P is not usually derived from one isolated hypothesis but from a set of statements T & A, containing the central laws of a theory (T) but also some auxiliary hypotheses and singular statements (A). If non-P results, some statement has to be given up since T & A & non-P is logically inconsistent. Which statement should be declared as false? This is Duhem's problem (which was later posed, in a similar form, by Quine).

The solution is systematic trial and error. If non-P obtains it is up to the researchers to speculate where the false assumption(s) is (are) located. They have to decide either to modify T or A and to replace the respective assumption by a new one. The result, T' & A, or T & A', is then again exposed to test in order

to find out whether T' & A, or perhaps T & A', is empirically more successful than T & A. This procedure is continued until the system can be confirmed, or the whole theory T is given up in favor of a new approach. However, deductivism does not specify when exactly a theory should be given up. There is no criterion of final falsification.

Deductivism can be conceived of as a set of rules. The rules of deductive logic are part of them, but they are not sufficient for empirical research. Deductivism does not need rules of induction. But it needs, in addition to deductive logic, *methodological* rules, which recommend what to do (or not to do) in various problem situations, e.g., when a theory is contradicted by empirical evidence. Most of them have the character of *heuristic principles* (Albert 1999, chapter II). They recommend thinking in a certain direction but do not exactly prescribe what to do. Furthermore, these rules are based on epistemological assumptions. Methodological rules are proposed because one is convinced that following them contributes to the aim of science, say, to finding true theories of high explanatory power. Of course, such rules, and the underlying assumptions, can themselves be critically discussed.

Deductivism was developed as a methodology of the natural sciences. However, the fundamental idea of deductivism can be applied to all empirical sciences, and to problem solving in general. Of course, theories as well as methods are quite different in the various sciences. The theory to be tested may be a system of universal statements, expressed as mathematical formulars, or it may be a set of singular statements about a historical event and its cause. The testing method may be a physical experiment, or a statistical correlation analysis in the social sciences. Yet the logic of research is the same. And the same kind of information processing takes place in ordinary learning. People always have beliefs and expectations, and when new information contradicts these beliefs, they sometimes correct them. The main difference between science and hypotheses testing in ordinary life is that in science hypotheses or theories are more elaborated and precise, and the methods of testing are more refined.

To repeat the main points, and to remove possible misunderstandings, let us consider some well-known objections and the answers given by deductivism.

Objection 1: Deductivism says that theories must be falsifiable, and that scientists should try to falsify them. However, scientific theories cannot be conclusively falsified.
Answer 1: Deductivism recommends to test theories severely, and not to defend them dogmatically. According to deductivism, conclusive falsifications are neither possible nor necessary.

Objection 2: Scientists do not normally try to falsify theories. Falsificationism is incompatible with the way famous scientists, like Galilei and Newton, proceeded.
Answer 2: First, deductivism gives methodological recommendations, not descriptions of how scientists actually proceed. Second, Andersson (1994) and Musgrave (1999, chapter 11) studied the examples from the history of science

pointed out by Kuhn and Feyerabend, and came to the conclusion that the historical facts are quite consistent with the rules of deductivism. It is true, for example, that Galilei questioned auxiliary hypotheses instead of the theory's central laws, but this is in accordance with deductivism.

Objection 3: Deductivism is insufficient as a methodology. The methods of science cannot be reduced to deductive logic.
Answer 3: Deductivism must not be equated with deductive logic. Deductive rules, such as Modus tollens, are a central part of deductivism, but deductivism contains, in addition, *methodological* rules, which are not logical laws or inference rules.

Objection 4: Deductivism cannot provide a logical proof that following its rules leads to progress, e.g., to achieving at true theories.
Answer 4: The idea that proceeding in accordance with deductivism leads to progress is itself a *hypothesis*. We could call it a methodological or epistemological hypothesis. Deductivism does not assume that this hypothesis can be deduced from the laws of deductive logic. But it can be critically discussed. And it can be demonstrated that deductivism solves more problems than other methodologies, especially, than inductivism.

6. Varied Testing and Specific Problems of Empirical Research

I have argued that for deductivism no problem of external validity arises. Nothing has to be generalized. The aspect of generality is introduced from the beginning, previously to collecting data, by formulating a universal theory. And the idea of generalizing to a more or less comprehensive population is, in deductivism, accounted for by giving the theory one starts with a higher or lower *degree of universality*. The higher the degree of universality of a statement, the higher is its empirical content, and its degree of falsifiability. By choosing a high degree of universality, we give us the chance of confirming a theory of high empirical content. But we also take a high risk that T will become falsified.

However, does this kind of reasoning really solve all problems associated with the ideas of generalization and representativeness? Consider the following questions. Does the empirical result we have found in rats also hold for humans? Isn't it problematic that many psychological theories have only been tested on American students of psychology? Would people behave in the same way in this experiment if they played with real money instead of play money? Undoubtedly, these are sensible questions, and it may seem that answering them requires some kind of generalization or inference by analogy. However, I now want to show that they can be well answered on the basis of deductivism, without recourse to external validity and inductive generalization.

It seems there are three types of questions in the above examples. There is, first, the question of *variety* in empirical studies. Variety is said to contribute to

external validity. Is variety also needed for theory testing? Second, we have the problem of transferring results gained in one population (e.g., rats) to another population (e.g., humans). This is said to require a conclusion by analogy. How does deductivism account for that problem? A third problem is associated with the distinction between experiments and field studies, or between artificial and natural conditions.

Evidently, *variety* is important in theory testing. Suppose an economic theory T says something about all humans. T is tested in a laboratory situation with students of economics, and is confirmed. But we cannot be sure that T is true. T could become falsified if we test it with people who are not students of economics, or in a more natural situation. Though T has been confirmed, there are infinitely many conditions under which T could become falsified. In order to test T thoroughly, we have to vary the test conditions. We should test T with different kinds of people, and in different kinds of situations. Proceeding like this puts T at a risk. If T passes all these varied tests, this contributes much to its confirmation. The test conditions need not be chosen randomly. It is much better to ask specific questions: In the light of the available knowledge, and from the perspective of rival theories, under what conditions should we expect T to become falsified? We might, e.g., come to the conclusion that T is unlikely to hold for subjects who are more experienced with the tasks they are given. We should then study this kind of situation since the result of such a study gives us much new information.

The second problem is how results gained with a certain population can tell us something about another population. Guala (2005, 195f.) proposes to use inferences by analogy. He illustrates this with an example from biomedical research. The inference, Guala says, here takes this form:

1. Humans have symptoms (disease) Y.
2. Laboratory animals have symptoms (disease) Y.
3. In laboratory animals, the symptoms (disease) are caused by factor (virus, bacteria, toxin, deficiency) X.
4. The human disease is therefore also caused by X.

Guala calls this an *analogical inference*, or an *external validity inference*. One may ask what is *valid* in such an inference. Guala concedes that analogical reasoning, like all inductive inferences, may lead to error. Yet he defends such reasoning.

I think, however, that calling this kind of research strategy an 'inference' which might be 'valid' is misleading. What scientists use here is simply a *heuristic principle* in the search for causes and explanations: 'In order to find the cause of Y in systems of type M, look what causes Y in other systems similar to M.' Heuristic principles help to develop ideas, but do not guarantee that these ideas are true, or probably true to some specified degree. This is quite clear in the above example. Take the hypothesis H, 'The human disease is caused by X.' What is it that could justify H, or to accept it as true? In biomedical research, the decision to accept such a hypothesis as true is never based on mere analog-

ical reasoning. It is never argued, e.g., Y is caused by X in rats, therefore it is justified to believe that Y is caused by X in humans. By contrast, it is insisted that the hypothesis has to be tested on humans, after it has been confirmed in rats. The step that decides whether H is acceptable or not is the deductive test of H, not some non-deductive inference. Hence, the logic of such research is *subsequent deductive testing of hypotheses*. Hypotheses are first tested on animals that are evolutionary related with man, and then on humans.

Third, why may it be problematic to investigate some phenomena only in laboratory experiments? Why are field studies useful, or even necessary, in social research? Why is it sometimes necessary to study human thought and action under 'natural' conditions? From an inductive perspective, one could argue that field results allowed analogical inferences (external validity inferences) to human behavior in everyday life, while experimental results did not allow such inferences. But if we reject inductive reasoning, what is the use of field studies? Why should it be relevant for theory testing whether the situation is artificial or natural, provided it is part of the theory's domain, which is the presupposition for any test? After all, control of extraneous variables is easier in experiments than in field studies.

On the basis of deductivism, there are several arguments for field research. First, it may be difficult to realize in a laboratory experiment the boundary and initial conditions of the theory in question. A theory T says that under certain conditions C something is the case. In the social sciences, C may include properties of the situation, of individual persons, of groups, or markets. Any test of T presupposes that C is realized. For example, if ego involvement is part of C but not realized in an experiment, this experiment does not count as a test of T. Many experiments have been criticized for being not sufficiently realistic. The subjects were not really engaged and interested in the result of what they do. (However, some experiments on emotions and social behavior were actually criticized with ethical arguments because the researchers caused in the subjects too much involvement, e.g., too strong emotions of fear, shame, or failure.) Another objection concerns *reactivity*, that is, the subjects' disposition to change their behavior merely because of being aware that they are studied.

Note, however, that these arguments only prove that field research *may* sometimes be necessary to test a theory in some part of its domain. In other cases, it may be the other way round. Scientific theories have many implications about events and processes that do not or very seldom occur in the natural world. This is evident in physics but holds as well in biology and economics. Such phenomena can only be studied in experiments, that is, in an artificial situation.

There is a further argument for field studies, which is particularly relevant in applied science. Suppose T says that, *ceteris paribus*, an increase in X leads to an increase in Y. T is confirmed in a series of experiments. Now we want to know whether the causal relation between X and Y also holds in a more natural situation of type N, e.g., in real markets, or in the natural environment of animals etc. In the experimental test, all other factors than X were held con-

stant. In the natural situations, such other factors do not remain constant, and affect Y, too. What will happen then? This cannot be predicted with T alone. We would need a comprehensive theory about all causal influences on Y, and we would, in addition, need methods to measure all this influences. In many cases, such a theory is not available. So one must find out empirically what happens in N where X and many other influences affect Y jointly. Such a study must be carried out in the natural situation one is interested in. The new hypothesis to be tested is, 'Under conditions N an increase in X leads to an increase in Y.'

7. The New Inductivism

Many social scientists who stress external validity and, recently, some experimental economists argue that social and economic research needs inductive reasoning instead of, or in addition to, deductive testing. What exactly do they mean by 'deductive' and 'inductive', and how do they argue for their view? Some examples may be helpful.

As a classic example for deductivism, Bardsley et al. (2010, 148) cite the test of Einstein's theory of general relativity by Eddington. It was predicted with the help of Einstein's theory that light passing near the sun is deflected by a specific amount. Newton's theory predicted only half of that amount. It was difficult to test this prediction, but it was possible during a solar eclipse in 1919. Einstein's theory was confirmed.

As an example from economics, Bardsley et al. present Chamberlin's (1948) experimental investigation of the determination of market prices. Here the theory (standard economic theory) was not confirmed: The subjects in the experimental market traded quantities at lower prices than was predicted.

According to Guala (2005, 48), the experiments on public goods are a typical case of theory testing, at least in the beginning. Here, too, standard economic theory was tested, and the results deviated somewhat from the prediction. The theory predicts that people would not contribute to the public project in the experiment, but actually they do, though, in repeated games, the initial level of contribution diminishes.

When Guala and Bardsley et al. describe hypothetico-deductive thinking, they have in mind not low-level hypotheses but major theories in the natural sciences, or standard economic theory. Instead of, or additionally to, testing major theories, they propose another kind of research, which they call inductive. As a typical example, Bardsley et al. cite Marmot's (2004) studies of health and social status. In his longitudinal studies of heart disease in British civil servants, Marmot found a 'social gradient': The mortality rate for the lowest of the four civil service grades was much higher than the mortality rate for the highest grade. Low occupational status was negatively correlated with health. How can this be explained? Marmot controlled for factors such as smoking, high blood pressure, and high blood sugar level. These factors were known to influence heart disease, yet they could explain only part of the correlation between status

and health. Marmot then found similar social gradients in other populations, including non-human primates. Higher-ranking laboratory monkeys were less susceptible to heart disease. Marmot reasoned that the decisive factor is stress. Stress is a cause of heart disease and diabetes, and is negatively correlated with social status.

As examples of inductive research in the social sciences and in economics, Bardsley et al. refer to Bryan and Test (1967) and Ball et al. (2001).

What exactly is the *inductive* element in such research? The new inductivists do not think that there are formal rules that lead from data to theories. They also know and emphasize that concepts and hypotheses always precede observation and collecting data. Induction does not mean blind search in the hope of stumbling on regularities. Yet it seems to them that there is a fundamental difference between Eddington's test of Einstein's theory and research like Marmot's. The hypotheses Marmot develops are *low-level* hypotheses. They do not provide very deep explanations, and they only explain a local phenomenon. "[. . .] Marmot's hypothesis about stress is not at all like the heroic hypotheses that provide classic examples for discussions of the hypothetico-deductive method." (Bardsley et al. 2010, 148)

Marmot tested statistical hypotheses which are not usually called 'theories'. "Our point is that while null and alternative hypotheses must have precise specifications, neither needs to be supported by any theory. There need be no bold conjecture to put to a severe test. Inductive enquiry can use formal statistical methods while investigating the flimsiest of hunches, expressed in the vaguest of terms." (Bardsley et al. 2010, 149)

Guala explains the difference between hypothetico-deductive and inductive research with respect to experimental work on public goods. Initially, public goods experiments were performed in order to test the standard theory. However, in subsequent experiments, the emphasis shifted. Scientists started to check the relevance of small variations in the experimental conditions. They varied features of the sample (male and female, economics and noneconomics students, British and Italian), of incentives, and of information. To be sure, such variations were guided by hypotheses. "But such hypotheses often take the form of rough insights, informal guesses that are not grounded in any well-structured and rigorously formulated theoretical system and are not in the corpus of economic theory." (Guala 2005, 48)

According to Bardsley et al. (2010, 27), such studies are attempts to "sharpen" empirical regularities. They aim at *regularity refinement*, and at testing the *robustness* of results. Are regularities robust to, say, cultural context, or task formulation?

In philosophy of science, induction was mainly discussed, and criticized, as a method of justification. Does inductive justification also play a role in new inductivism? We have already analyzed inferences by analogy, as proposed by Guala, and external validity inferences, which are said to be of central importance in inductive research. Furthermore, Bardsley et al. refer to methods of statistical testing as part of inductive reasoning. "Take the proposition [. . .] that there is

less mortality in the higher civil service grades than in the lower ones. From the evidence of Marmot's sample, we can infer that this proposition is very probably true. The 'very probably' here is not a degree of belief, derived by updating prior probabilities; it is a statement about the frequency with which data of the kind observed would be generated by a particular null hypothesis." (Bardsley et al. 2010, 150) Obviously, the authors restrain from Bayesianism. And they seem to assume that statistical testing allows judgments about the probability of hypotheses.

The main theses and arguments of new inductivism in experimental economics that I now want to discuss can be summarized as follows.

1. Experimental economics needs *low-level theorizing* and *low-level hypothesis testing*. 'Low-level' here means theorizing on a level that is near the observational level. Such research contributes to *regularity refinement*, and to testing regularities for *robustness*.
2. Low-level research is not accounted for by deductivism. Therefore, and for further reasons, deductivism is unconvincing, or at least not sufficient, as a methodology of experimental economics. An *inductive* methodology is more adequate.
3. Inductive reasoning includes as a central part *external validity inferences* and statistical methods, by which hypotheses can (sometimes) be proven to be *very probably true*.

I fully agree with thesis (1). Of course low-level work is necessary. There are always facts that call for an explanation but cannot currently be explained by major theories. In such cases one tries explanations with low-level hypotheses, which may later on themselves be explained by deeper assumptions. Furthermore, it is important to test experimental results and low level hypotheses for their robustness. It is necessary to repeat experiments, and to vary the conditions. It may be true that in experimental economics low-level work was neglected for a long time and has now been identified as necessary and important.

However, I do not agree with theses (2) and (3). In contrast to what inductivists claim, deductivism can well account for low-level work in the experimental sciences. Such work is necessary for testing major theories. Major theories are not usually falsified or confirmed by isolated observations. Popper (1959, 66) stressed that "non-reproducible single occurrences are of no significance to science. Thus a few stray basic statements contradicting a theory will hardly induce us to reject it as falsified. We shall take it as falsified only if we discover a *reproducible effect* which refutes the theory. In other words, we only accept the falsification if a low-level empirical hypothesis which describes such an effect is proposed and corroborated. This kind of hypothesis may be called a *falsifying hypothesis*." Thus, according to deductivism, research requires work on different levels, including the level of observation and experiment.

Why do Guala and Bardsley et al. call such work inductive? They admit that low-level research is also guided by hypotheses. They also concede that there are no formal (inductive) rules for generating theories from observations.

What then is the 'inductive' aspect in low-level work that distinguishes it from hypothetico-deductive research? It seems to be nothing else than the degree of universality and explanatory power of the hypotheses involved. When they are part of a major theory, research is called deductive. When the hypotheses refer to things and properties near the observational level, research is instead named inductive.

It seems that inductivists have a rather restricted view of deductivism. They think deductivism is fixed on major ('heroic') theories, has no place for low-level hypotheses, and hence cannot account for low-level work. But this is a misunderstanding. The methodology of deductive testing is by no means restricted to hypotheses or theories of a certain kind. Einstein's theory was deductively tested. But singular historical statements can be deductively tested as well. Even text interpretation can be conceived of as deductive testing.

As to (3), I have already criticized external validity and analogical inferences. Bardsley et al. claim, in addition, that hypotheses like Marmot's could be shown to be 'very probably true'. This is indeed an inductive idea. Induction as a method of proof means an inference from observation statements to a hypothesis by which the hypothesis is shown to be true with certainty, or true with a certain probability. Unfortunately, Bardsley et al. do not say how one can calculate probabilities of hypotheses or theories. Here is again what they say: "The 'very probably' here is [. . .] a statement about the frequency with which data of the kind observed would be generated by a particular null hypothesis." (2010, 150) I must confess that I don't understand what the authors here mean. As they say, when a null hypothesis is tested, the probability of data given the hypothesis is calculated. The hypothesis itself is then either rejected or not. Its probability cannot be calculated by this procedure. Yet the authors suggest that statistical testing somehow gives us probabilities of statistical hypotheses. This is false. Even less is it possible to calculate a probability for a scientific theory from which a null hypothesis (or alternative hypothesis) is deduced.

Traditionally, methodology is a rather philosophical enterprise, which has therefore been widely ignored by scientists. The new inductivism attempts to bring methodology in touch with the practice of social science and economics, and to render it more realistic and adequate. For this purpose, Guala and Bardsley et al. took up ideas from the textbooks of social research which are not found in general philosophy of science, such as internal and external validity.

In principle, I welcome the project of bringing general methodology into contact with scientific practice. However, the way this is done here is not convincing. The new inductivists suggest that deductivism is problematic and insufficient as a methodology of science. But they describe deductivism inadequately, namely, as a view that only applies to the major theories of science. As an alternative, they propose a new form of inductive thinking. But they have not been able to demonstrate how the well-known problems of inductive inferences could be solved.

I have tried to show that deductivism, adequately understood, can solve all problems of experimental research associated with the idea of external valid-

ity. If this is true, there is no reason to replace deductivism by inductive ideas, or to complement it with such ideas, which would introduce into experimental economics a lot of confusion and unsolved problems.

References

Albert, H. (1968), *Traktat über kritische Vernunft*, Tübingen: Mohr Siebeck (English translation: *Treatise on Critical Reason*, Princeton 1985).

— (1999), *Between Social Science, Religion and Politics*, Amsterdam: Rodopi.

Albert, M. (2003), "Bayesian Rationality and Decision Making: A Critical Review", *Analyse & Kritik* 25, 101–117.

Andersson, G. (1994), *Criticism and the History of Science*, Leiden: Brill.

Ball, S., C. Eckel, P. J. Grossman and W. Zame (2001), "Status in Markets", *Quarterly Journal of Economics* 116, 161–188.

Bardsley, N., R. Cubitt, G. Loomes, P. Moffatt, C. Starmer and R. Sugden (2010), *Experimental Economics: Rethinking the Rules*, Princeton: University Press.

Bryan, J. H. and M. A. Test (1967), "Models and Helping: Naturalistic Studies in Aiding", *Journal of Personality and Social Psychology* 6, 400–407.

Campbell, D. T. (1957), "Factors Relevant to the Validity of Experiments in Social Settings", *Psychological Bulletin* 54, 297–312.

— and J. C. Stanley (1963), "Experimental and Quasi-Experimental Designs for Research on Teaching", in: Gage, N. L. (ed.), *Handbook of Research on Teaching*, Chicago: Rand McNally, 171–246.

Chalmers, A. F. (1999), *What Is This Thing Called Science?*, Indianapolis: Hackett Publishing Company.

Chamberlin, E. (1948), "An Experimental Imperfect Market", *Journal of Political Economy* 56, 95–108.

Feyerabend, P. (1975), *Against Method*, London: New Left Books.

Gadenne, V. (2002), "Hat der kritische Rationalismus noch etwas zu lehren?", in Böhm, J. M, H. Holweg and C. Hoock (eds.), *Karl Poppers kritischer Rationalismus heute*, Tübingen: Mohr Siebeck, 58–78.

— (1976), *Die Gültigkeit psychologischer Untersuchungen*, Stuttgart: Kohlhammer.

Guala, F. (2005), *The Methodology of Experimental Economics*, Cambridge: University Press.

Howson, C. and P. Urbach (1989), *Scientific Reasoning: The Bayesian Approach*, La Salle: Open Court.

Kuhn, T. S. (1962), *The Structure of Scientific Revolutions*, Chicago: University Press.

Marmot, M. (2004), *Status Syndrom: How Your Social Standing Directly Affects Your Health and Life Expectancy*, London: Bloomsbury.

Miller, D. (1994), *Critical Rationalism: A Restatement and Defence*, Chicago: Open Court.

Musgrave, A. (1999), *Essays on Realism and Rationalism*, Amsterdam: Rodopi.

Popper, K. (1959), *The Logic of Scientific Discovery*, London: Hutchinson (English translation of *Logik der Forschung*).

— (1994[1934]), *Logik der Forschung*, 10th ed., Tübingen: Mohr Siebeck.