Søberg, Morten

Working Paper

The Duhem-Quine thesis and experimental economics. A reinterpretation

Discussion Papers, No. 329

Provided in Cooperation with:
Research Department, Statistics Norway, Oslo

Suggested Citation: Søberg, Morten (2002) : The Duhem-Quine thesis and experimental economics. A reinterpretation, Discussion Papers, No. 329, Statistics Norway, Research Department, Oslo

This Version is available at:
https://hdl.handle.net/10419/192311

Discussion Papers No. 329, August 2002
Statistics Norway, Research Department

Morten Søberg

The Duhem-Quine thesis and experimental economics
A reinterpretation

Abstract:
The Duhem-Quine thesis asserts that any empirical evaluation of a theory is in fact a composite test of several interconnected hypotheses. Recalcitrant evidence signals falsity within the conjunction of hypotheses, but logic alone cannot pinpoint the individual element(s) inside the theoretical cluster responsible for a false prediction. This paper considers the relevance of the Duhem-Quine thesis for experimental economics. A starting point is to detail how laboratory evaluations of economic hypotheses constitute composite tests. Another aim is to scrutinize the strategy of conducting a series of experiments in order to hem in the source(s) of disconfirmative evidence. A Bayesian approach is employed to argue that reproducing experiments is not necessarily useful in terms of identifying correct causes of recalcitrant data.

Keywords: Experimental economics, methodology, Duhem-Quine thesis.

JEL classification: B41, C90

Acknowledgement: Constructive comments and helpful suggestions from Francesco Guala, Aanund Hylland, Chris Starmer, Vernon Smith and seminar participants at the Economic Science Association's meeting in Barcelona and the University of Toulouse are highly appreciated. Funding from the Norwegian Research Council is gratefully acknowledged. The conventional disclaimer applies.

Address: Morten Søberg, Statistics Norway, Research Department. E-mail: morten.soberg@ssb.no

Discussion Papers	comprise research papers intended for international journals or books. As a preprint a Discussion Paper can be longer and more elaborate than a standard journal article by including intermediate calculation and background material etc.

Abstracts with downloadable PDF files of
Discussion Papers are available on the Internet: http://www.ssb.no

For printed Discussion Papers contact:

Statistics Norway
Sales- and subscription service
N-2225 Kongsvinger

Telephone: +47 62 88 55 00
Telefax: +47 62 88 55 95
E-mail: Salg-abonnement@ssb.no

1. Introduction

The Duhem-Quine thesis refers to the relationship between theory and evidence. In particular, the thesis posits the non-separability and the non-falsifiability of single theoretical hypotheses (Cross, 1998). Non-separability means that empirical predictions can only be deduced from clusters of interconnected hypotheses. Anomalous evidence consequently implies falsity somewhere inside a theoretical network. In such cases non-falsifiability obtains because pure logic alone cannot pinpoint the exact culprit(s) in a theoretical maze responsible for a false prediction. Thus, even if the evidence is indisputable, the thesis essentially questions whether any empirical assessment of theory can be conclusive.

This paper addresses the significance of the Duhem-Quine thesis for laboratory experimentation in economics[1]. Its broad objective is to clarify the content of the thesis and to reinterpret its relevance in experimental economics. The number of previous contributions to this discussion is surprisingly low. More importantly they contain somewhat contrasting accounts of how the thesis applies and the degree to which it reflects inherently problematic features of economic experiments. On the one hand, Starmer (1999) believes that the Duhem-Quine issues have been both downplayed for strategic reasons and inadequately considered. He posits that the pioneering experimentalists - predominantly Vernon Smith and Charles Plott - have, first and foremost, been concerned with establishing experimental economics as an acceptable branch of empirical enquiry:

"At the time they were writing it seems reasonable to suppose that the prevailing view was a skeptical one: that economics is a non-laboratory science. As such, it would be understandable if those promoting experimental methods saw themselves fighting a tough corner in which a robust defense was called for. If so, the incentives were there to be up-beat about the benefits from experiments while playing down the difficulties of interpretation" (1999, p. 22).

Starmer proceeds to highlight experimental control and the domain of economic theories as two crucial Duhem-Quine issues in experimental economics. A vital reference is made to Vernon Smith's (1982) seminal paper, which considered how a controlled laboratory environment necessitates a satisfactory linkage between monetary rewards and subjects' behavior in the laboratory. For his part, Starmer argues that any claim to experimental control amounts to an ever-present auxiliary hypothesis, which he considers cannot easily be singled out from a main theory and tested in isolation. Moreover,

[1] Cross (1982), Heijdra and Lowenberg (1986), and Sawyer et al. (1997) elaborate on the Duhem-Quine thesis with regard to non-experimental, empirical evaluations of economic theories.

Starmer holds that any experimental test is unavoidably non-separable from the auxiliary hypothesis that the theory in question applies to the experimental context. In the case in which a theory is domain-specific, an experimental evaluation conducted in an inappropriate context will not yield any meaningful assessment of the empirical validity of this theory.

Then again, Vernon Smith's (1994) paper was the first study that included an explicit analysis of Duhem-Quine issues in relation to economic laboratory data. Although many experimental economists may not refer openly to the Duhem-Quine thesis, Smith considers their actual conduct to be vastly inspired by its intrinsic logic: When theories are poorly supported by data, the experimentalists seek to establish an anatomy of failure. Accordingly, instructions are investigated for clarity and new experimental trials are run with different subjects, alternative financial incentive structures and so forth. Smith's key stance is that even though a host of Duhem-Quine issues regularly arise in the laboratory, they can to a large extent be resolved by recourse to further experiments.

The detailed purpose of this paper is to, first, build on Starmer's argument and detail how non-separability applies to experimental economics. This will be achieved by characterizing three types of auxiliary assumptions deemed indispensable when deriving observable predictions/benchmarks against which laboratory data can be assessed and interpreted. The second aim is to investigate the validity of Smith's position that sequences of experiments can be conducted in order to alleviate the problem of non-falsifiability. A useful stepping-stone for such an analysis is to spell out the separate components of the merged Duhem-Quine thesis. This distinction is employed to argue that only an explicitly Quinean perspective justifies running interrelated experiments designed to locate sources of error. The upshot of this viewpoint is that it may generally favor saving a main hypothesis at the expense of refuting auxiliary hypotheses, regardless of the number of reproduced experiments.

The remainder of the paper is organized relative to the listed intentions. Hence, the next section provides a detailed introduction to Duhem and Quine's theses. The subsequent section contains a discussion of non-separability in experimental economics, and the fourth section is devoted to the issue of non-falsifiability. The last section concludes.

2. Duhem's and Quine's theses

The Duhem-Quine thesis can be traced to the physicist Pierre Duhem (1861-1916) and the philosopher Willard Van Orman Quine (1908-2000), and their respective works "The Aim and Structure of Physical Theory" (1906) and "Two Dogmas of Empiricism" (1951). As his title indicates, Duhem's

examination was confined to experimental physics, and was explicitly demarcated from other natural sciences such as chemistry and physiology[2]. Quine's contribution to the Duhem-Quine thesis was developed in the context of a philosophical discussion of two empiricist dogmas: The asserted distinction between analytic and synthetic statements as well as reductionism, that is, the belief that each statement taken in isolation can admit of disconfirmation. Quine superseded Duhem's restricted scope and sought to apply his position to the "totality of our so-called knowledge or beliefs, from the most casual matters of geography and history to the profoundest laws of atomic physics or even of pure mathematics and logic" (Quine, 1980, p. 42). Below, the individual positions of Duhem and Quine are elaborated on separately (alternative expositions of the theses can be found in Vuillemin, 1986 and Gilles, 1993, Chapter 5).

2.1. The Duhem thesis

Duhem's primary contention is non-separability. This is the assertion that theoretical statements cannot, singly, be disconfirmed. The reason is that hypotheses in physics do not by themselves imply any observational consequences. Instead, predictions are at all times deducted from conjunctions of theoretical hypotheses, auxiliary hypotheses and background knowledge. In addition, non-separability pertains to the laboratory data, since the evidence generated in a physics experiment may, in the main, not allow for direct sensory observation. Therefore, both the construction of empirical data and the interpretation of these in practice require sophisticated instruments, the design and operation of which rely upon other theories as well[3].

For Duhem, non-separability is an empirical claim and does not apply *a priori* to all sciences. Thus, the validity of non-separability is context-dependent and needs to be ascertained on a case-by-case basis. However, if non-separability does obtain, then non-falsifiability is implied (Ariew, 1984, p. 318). In Duhem's words, so-called crucial experiments are prohibited:

"In sum, the physicist can never subject an isolated hypothesis to experimental test, but only a whole group of hypotheses; when the experiment is in disagreement with his predictions, what he learns is that at least one of his hypotheses constituting this group is unacceptable and ought to be modified; but the experiment does not designate which one should be changed" (Duhem, 1991, p. 187).

---

[2] Duhem distinguishes between experiments of application and experiments of testing. Application experiments deal with practical problems and draw extensively upon well-corroborated theories. On the other hand, test experiments, which are the chief concern of Duhem, aim at discovering whether theories are empirically accurate or not (cf. Duhem, 1991, p. 183).

[3] Implicit in this reading of Duhem is the view of an experiment as a theory-driven procedure: Theoretical hypotheses steer the experimental process in terms of formulating questions and guiding the experimental design, and the laboratory evidence is interpreted relative to the underlying theory.

Equivalently, Duhem asserts the impossibility of a direct disproof of an individual theoretical hypothesis. The argument can be illuminated by resort to deductive logic. Consider a conjunction of a main hypothesis *h*, an auxiliary hypothesis *a* and background knowledge *b* that implies the empirical consequence *e*:

$$(1)\, h \wedge a \wedge b \rightarrow e$$

where $\wedge$ denotes the logical operator "and", whereas the arrow $\rightarrow$ signifies "implies".

Non-separability is taken to mean that *e* cannot be deduced from *h* by itself: The main hypothesis entails observational consequences only when conjoined with relevant auxiliary assumptions and background knowledge. Now assume that the predicted observational statement does not obtain so that *e* is untrue. A straightforward modus tollens argument then implies that at least one of the premises constituting the explanans of *e* is at fault: *h* or *a* or *b*, or some combination thereof (Jeffrey, 1989):

$$(2) \neg e \rightarrow \neg(h \wedge a \wedge b) \leftrightarrow \neg h \vee \neg a \vee \neg b$$

Here $\neg$ represents "not", $\vee$ means the logical operator "or", and $\leftrightarrow$ asserts equivalence.

Duhemian non-falsifiability is, basically, the problem that pure logic cannot pinpoint the exact location of the false prediction's hit. When empirical consequences deduced from these premises conflict with observational facts, it is impossible to conclude on logical grounds which of the premises is at fault and ought to be abandoned. A corollary argument of Duhem's is that non-falsification of a theoretical network does not imply its verification: The possibility of future recalcitrant empirical data remains, as does the discovery of unknown explanations more satisfactory than the present alternatives (Laudan, 1976, p. 157). Importantly, what drives Duhem's analysis of non-separability and non-falsifiability is an organic, as opposed to mechanic, perception of physics:

"Physics is not a machine that can be taken apart; one cannot try each piece in isolation and wait, in order to adjust it, until its solidity has been checked carefully. Physical science is a system that must be taken as a whole; it is an organism in which one part cannot be made to function without the more remote parts coming into play, some more than others, but all to some degree. ... The watchmaker to whom one gives a watch that does not function separates all the wheels and examines them one by one until he finds the one which is defective or broken. The doctor to whom a patient appears cannot dissect him in order to establish his diagnosis; he has to guess at the seat and cause of the ailment solely by inspecting disorders affecting the whole body. The physicist concerned with remedying a defective theory resembles the doctor and not the watchmaker" (1991, pp. 187-8).

Duhem argues that the scientist needs to resort to "good sense" or sagacity as a guidance principle when deciding upon how to deal with disconfirmative experimental evidence. By good sense is meant an extra-logical faculty; something resembling intuition as opposed to deduction, *raison* rather than *raisonnement* (Ariew, 1984, p. 317):

"These motives which do not proceed from logic and yet direct our choices, these 'reasons which reason does not know' and which speak to the ample 'mind of finesse' but not to the 'geometric mind', constitute what is appropriately called good sense" (1991, p. 217).

In situations where theoretical hypotheses have been struck by experimental evidence there may, in many instances, be good reasons for proceeding in certain ways, be it a resolve to abandon a main hypothesis or a choice to refute an auxiliary one. The main implication of Duhem's thesis is that such decisions cannot be conclusive.

2.2. The Quine thesis

The Quine thesis is grounded by an explicit reference to the Duhemian notion of non-separability and declares that:

"... our statements about the external world face the tribunal of sense experience not individually but only as a corporate body" (Quine, 1980, p. 41).

However, while Duhem confines his focus to experimental tests of theoretical physics, Quine respects no such boundary. Instead, he envisages the set of conjoined statements about the world as extending and ramifying until they include the totality of science. Quine conceives of a "field of force" as an impressionistic description of science. This is an entity constrained only by the whole of human knowledge, including myths, fiction, and logical categories. Visually, the field of force may be perceived with an "interior", say, a hard core, that is more resistant to modification than its "periphery" (also see Lakatos, 1978). The latter is defined as degrees of closeness to sensory events; experience. For instance, Quine posits that some statements have a sharper empirical reference and comparatively greater vulnerability *vis-à-vis* recalcitrant data than

"... highly theoretical statements of physics or logic or ontology. The latter statements may be thought of as relatively centrally located within the total network, meaning merely that little preferential connection with any particular sense data obtrudes itself" (1980, p. 44).

Statements that are located centrally within the system transcend all empirical data in that they have no straightforward descriptive import. Remoteness from experience does not imply independence from

experimental test, but in practice Quine contends that disconfirmative empirical evidence ought to bear on the statements located closest to the empirical periphery:

"The edge of the system must be kept squared with experience; the rest, with all its elaborate myths or fictions, has as its objective the simplicity of laws" (1980, p. 45).

Hence, Quine argues that adjustments of the union of statements heed the principle of Ockham's razor: The maximum of minimum mutilation. Equivalently, disturbances of the system are to be minimized. This entails granting central elements of the network, such as mathematics and logical principles, *de facto* immunity from the revision processes. In practice, the experimenter ought to defuse inconsistencies with evidence by sacrificing statements that disturb the existing theoretical statements the least:

"... the experimenter picks in advance the particular sentence that he will choose to sacrifice if the experiment refutes his compartment of theory. He will select the sentence with a view to disturbing the existing theory least, unless strong simplicity considerations intervene" (Quine, 1986, p. 621).

Another labeling of Quine's recommended revision strategy is that of tenacity, meaning that theoretical systems are resistant to dramatic changes in the face of recalcitrant empirical observations (Cross, 1998, p. 107). Thus, the Quine thesis implies that a main theory h should be retained or saved from refutation at the expense of adjusting an auxiliary hypothesis. However, the Quine thesis has sometimes been interpreted as a relatively stronger assertion, namely the claim that there always exists an appropriate auxiliary hypothesis a' which, when conjoined with a main theory, entails the empirical observation e:

$$(3)\,\forall h\,\forall b\,\forall e\,\exists a'\,(h \wedge a' \wedge b) \rightarrow e$$

where $\forall$ ($\exists$) is the universal (existential) quantifier.

Although a logical possibility, this reading has been criticized as implying that the trivial revision of auxiliary hypotheses is sufficient to rescue a particular theory from refutation by anomalous empirical evidence (Grünbaum, 1976, pp. 117-118). A related notion is termed underdetermination, i.e., the belief that there are an infinite number of theories which might fit specific empirical facts more or less adequately. The practical meaning of underdetermination is not clear-cut because actual scientific conduct is not usually characterized by the availability of an infinitely large number of alternative auxiliary hypotheses. Instead the relevant research context usually strips down the set of possible

alternatives (Nickles, 1988). Quine himself concedes as much when he notes that, in reality, empirical evidence affects clusters of theoretical statements that tend to be graded off (1980, p. viii).

3. Non-separability in experimental economics

The meaning of both Duhemian and Quinean non-separability is that auxiliary hypotheses and background knowledge are required in order to connect a theoretical hypothesis to an empirically assessable prediction. The aim of this section is to epitomize how non-separability, so defined, applies to laboratory experimentation in economics.

Following Smith (1982), the defining ingredients of an economic laboratory experiment are termed environment, institution, and behavior. By environment is meant a collection of assumptions that specify the number and kind of subjects, their preferences/technology and endowments, whereas an institution can be defined as a set of rules that specify the message space of the subjects, and how messages may be converted into binding contracts and allocations of resources. Finally, the empirical properties of an economic theory are assessed in terms of the outcomes determined by actions chosen by human subjects in the laboratory. The key point is that environmental and institutional specifications make up the context within which laboratory behavior is generated and observed.

However, a Duhem-Quinean reading of laboratory environments and institutions would be to consider them auxiliary assumptions which assist the derivation of empirical predictions from formal theory. The reason is that theoretical statements in economics usually consist of abstract formulations based on various definitions and concepts (see, e.g., Hausman, 1992). One example is expected utility maximization. As such, it can be considered an abstract and formal assumption regarding the behavior of human agents, but a mathematical definition of expected utility maximization is not equivalent to a statement about empirical data. Therefore, the chief purpose of an environment and institution is to establish an intended domain of the theory, that is, to provide an actual empirical reference of theoretical hypotheses. Consider a laboratory evaluation of expected utility maximization within an environment consisting of one or more buyers and institutional rules defined by an English auction[4]. This test is empirically operational because the conjunction of these hypotheses facilitates observable predictions in the form of bids to buy and confirmed auction prices.
Testing economic theory involves additional assumptions regarding the experimental design. These include the specification of the terms and labels to be included in the experimental instructions, the

[4] The English auction allows buyers to specify bids to buy that are communicated to all market participants. The object put up for sale is awarded to the highest bidder and the winner pays his bid price.

type of subjects and financial reward structure to be employed, and the number and the length of laboratory sessions. The latter specifications command particular importance since they determine the kind of data the experiment will generate and help pin down relevant empirical performance measures. Hence, they serve to narrow and concretize the empirical reference of the theory to be tested.

Table 1 summarizes the constituents of a laboratory test of economic theory as sketched above. Specifically, the auxiliary assumptions that join a theoretical statement (h) and its empirical implication (e) are organized in three separate categories: $a^{Environment}$, $a^{Institution}$ and a^{Design}. In addition, the outline includes background knowledge. It is defined as consisting of a relevant body of statistical procedures, theoretic and experimental facts that guide the objective of the test, the choice of terminology and symbolism, and the analysis of the laboratory data. Also, background knowledge may affect experimental tests by steering the specification of the various auxiliary assumptions, e.g., it contributes to the identification of the appropriate number of buyers, influences the choice of length of laboratory sessions, etc.

***Table 1*. Constituents of laboratory tests of economic theories**

Symbol	Description
h	Abstract theory and/or formal assumptions
$a^{Environment}$	Collection of agents' characteristics; number of subjects, their preferences and endowments
$a^{Institution}$	Rules that specify agents' message space, the dissemination of messages and how messages transform into contracts/allocations of resources
a^{Design}	Instructions, financial incentives, number/length of laboratory sessions, unit of observation
b	Statistical procedures, the whole of economic theory, the state of experimental knowledge
e	Predicted observable outcome; assessable empirical reference of theory

In conclusion, non-separability applies in the sense that an economic theoretical statement *per se* tends to be devoid of empirical reference, and a conjunction of theory, auxiliary assumptions and background knowledge is needed in order to derive an empirical prediction. However, this contention warrants two further comments.

First, the outlined argument has been restricted to the non-separability of laboratory tests of theoretical statements, whereas laboratory experimentation in economics clearly encompasses a wide range of qualitatively different procedures as well. Two alternative types of laboratory experimentation are theory stress-tests and search for empirical regularities (Davis and Holt, 1993). The former are classified as attempts to reveal the limits of applicability of economic propositions, whereas the latter

denote exploratory or heuristic experiments that can be largely independent of theoretically motivated research questions. Smith (1994) adds to this list experiments that seek to compare different laboratory environments and institutions. Yet other experiments are performed with a view to evaluating policy proposals and exploring the performance properties of theoretically unexplored institutional designs. A condensed taxonomy of the prevalent laboratory procedures in economics could involve distinguishing between theory-driven and data-driven experiments (also see Steinle, 1997). A theory-driven experiment incorporates a main theoretical hypothesis, which is conjoined with a set of auxiliary hypotheses to imply a prediction. In addition to well-elaborated abstract statements, the theoretical input may consist of tentative conjectures or hypotheses induced from empirical regularities. The data generated by such an experiment refer to, and are interpreted relative to, the derived prediction. In contrast, the data-driven category subsumes all types of experiments which are neither conceived nor designed under the strict supervision of a theory. However, any such data-driven experiment requires specifications in the form of environment, institution and design, which typically facilitate the calculation of benchmarks against which data may be compared. The main implication is that non-separability prevails, since the production of experimental evidence is a function of conjoined auxiliary hypotheses.

The second qualification is that the sketched-out description of non-separability refers to any singular experiment. Nonetheless, additional and related experiments, which differ from the original test in that a specific type of auxiliary hypothesis is varied, can, in principle, be conducted. Thus, the individual effect of an environmental auxiliary hypothesis - say, the number of buyers participating in an English auction - may be ascertained by means of running a series of experiments which differ with regard to the number of employed buyers. As a result, non-separability is arguably non-strict when evaluated relative to directed sequences of experiments. To paraphrase Duhem, in this dynamic sense an experimenter in economics may be more like a watchmaker than a doctor.

4. Non-falsifiability in experimental economics

Non-falsifiability is triggered by the generation of recalcitrant experimental evidence. To fix ideas, consider the case where a main hypothesis (h) conjoined with the auxiliary hypothesis (a) implies the empirical observation (e)[5].

[5] The phrase "main hypothesis" is not intended to be restrictive. It may denote either a tentative conjecture or a well-elaborated theoretical statement. Also, in the context of a data-driven experiment, h might be a specified environment, the auxiliary hypothesis a an institution and e an empirical benchmark.

$$(4)\; h \wedge a \rightarrow e$$

The union of *h* and *a* is falsified if the laboratory data contradict *e*. Therefore, non-falsifiability obtains in the sense that the experimenter is now faced with three competing and mutually exclusive explanations of the false prediction: Either *h* is true and *a* false, or *h* is false and *a* true, or both *h* and *a* are at fault and should be abandoned[6]:

$$(5)\; \neg e \rightarrow (h \wedge \neg a) \vee (\neg h \wedge a) \vee (\neg h \wedge \neg a)$$

Vernon Smith (1994) contends that the norm in the experimental economics community is to respond to disconfirmation by conducting series of interrelated experiments deliberately designed to isolate the source(s) of falsity[7]. The idea is that a well-designed sequence of experiments can assist the choice of a particular explanation for the recalcitrant evidence. Specifically, Smith argues that new experiments are conducted in order to check the validity of one or more auxiliary assumptions, for instance, hypotheses asserting clarity of instructions and adequacy of financial incentives. The objective of this part of the paper is to elaborate upon such experimental processes and to evaluate their effect on the Duhem-Quinean problem of non-falsifiability.

An essential characteristic of experimental methodology is the possibility of replicating and reproducing experiments. Replications of an experiment entail collecting additional laboratory data using similar conjunctions of environment, institution and design. More specifically, replication is successful if the new data do not differ in a statistically significant manner from the original results[8]. In contrast, a reproduction amounts to a variation of the original experiment, the usual objective of which is to investigate the robustness of the empirical phenomenon in question (Cartwright, 1991).

One reading of Smith is that experimental economists resort to reproducing experiments with the aim of weeding out the possibility that a random effect, which falsified the auxiliary hypothesis, has

---

[6] Background knowledge is omitted from the conjunction in equation (4) while just one type of auxiliary hypotheses is included. One justification for this simplification is that a realistic research context in experimental economics typically restricts the set of possible explanations of a false prediction. This point is made in greater detail below.

[7] A related issue is how purely disconfirmative findings are regarded amongst (experimental) economists. Roth (1994) claims to have observed a tendency among experimentalists to avoid reporting "negative" results altogether and is worried about the extent to which biased data selection, i.e. leaving out recalcitrant or "unfitting" data, plays a role in the analysis and interpretation of experimental data.

[8] Therefore, the meaning of replication in experimental economics and alternative empirical methods of analysis, such as econometrics, differs markedly. In the latter case it primarily refers to the ability of other investigators to repeat identical analyses based on the same data set (see, for instance, DeWald et al., 1986).

occurred. Zahar (1983, pp. 154-161) contains a concise analysis of this strategy for dealing with recalcitrant experimental evidence. His position entails running a series of reproductions characterized by non-trivial and relevant alterations of the auxiliary hypothesis. These experiments should be empirically equivalent in the sense that the main hypothesis *h* conjoined with any of the alternative auxiliary hypotheses a_i, $i \in \{1,2, .., n\}$, implies the same empirical observation *e*. Zahar then considers the case where each experimental reproduction produces recalcitrant evidence[9]. Such a sequence of experiments with an arbitrary stopping point is charted formally in equation (6). The intuitive idea is that a stream of consistently disconfirmative data indicates that the main hypothesis *h* is at fault and should be refuted.

$$
\begin{array}{ll}
 & h \wedge a_1 \rightarrow e, \neg e \\
 & h \wedge a_2 \rightarrow e, \neg e \\
(6) & \quad : \\
 & \quad : \\
 & h \wedge a_n \rightarrow e, \neg e
\end{array}
$$

The applicability of Zahar's procedure in experimental economics may be addressed by discussing, first, the requirement of non-trivial changes of auxiliary hypotheses and empirical equivalence accompanied by, second, an analysis of the type of inference that may be induced from consistent recalcitrant evidence.

Noticeably, any economic laboratory experiment amounts to a local context that strips down the battery of alternative auxiliary hypotheses. To illustrate, an experimental evaluation of the English auction trivially prohibits variations of institutional auxiliary hypotheses. Furthermore, practicalities in the form of time and budget constraints typically restrict the range and type of possible reproductions. Analogous considerations apply to the a^{Design} category of auxiliary hypotheses, which usually contain the assumptions that the choice of subject pool is irrelevant to the production of laboratory data and that the implemented financial reward structure is adequate to sufficiently motivate the employed subjects. Testing the validity of such auxiliary hypotheses could consist of experiments using students, teachers and professionals or, say, a two-fold, four-fold and eight-fold increase in financial rewards, respectively.

[9] Arguably, the Duhem-Quine problem of non-falsifiability arises only in the case of consistent recalcitrant evidence (Zahar, 1983, p. 155).

Importantly, variations in the number of agents, types of subjects and the size of potential payoffs generally do not affect the derived prediction (the calculated benchmark) of a theory-driven (data-driven) experiment. As a case in point, game theoretic predictions do not differ across different members of the socioeconomic population, and the benchmark pertaining to a market experiment typically is invariant with the size of available payoffs. In summary, non-trivial variations of an auxiliary hypothesis tends to be consistent with comparatively short series of reproductions, whereas empirical adequacy may not be an exceedingly restrictive requirement.

The second and comparatively more urgent issue concerns the interpretation of streams of recalcitrant data. Zahar's argument is that if the main hypothesis is to be rescued, i.e. held to be true, the falsity of each of the considered auxiliary hypotheses has to be assumed. Consequently, the n independent experiments yields an $n + 1$ tuple of truth-values {true, false, false, .. , false} for the set of hypotheses $\{h, a_1, a_2, .. , a_n\}$. The probability that the union of auxiliary hypotheses is false equals $(½)^n$, which converges (rapidly) to zero as the number of experiments grows large.

One problem with this quasi-inductive solution to Duhem-Quinean non-falsifiability is the assumption that h and a_i are stochastically independent. This is equivalent to assuming that the probability of h is independent of whether a_i is true or false[10]. It is generally highly questionable whether this assumption may apply to main and auxiliary hypotheses that have been part of the same research program for some time (Strevens, 2001). A second problem is that Zahar's argument basically responds to a diachronic issue, that is, it relates to how the probability of one specific explanation is affected by consistent recalcitrant experimental evidence. It does not provide any synchronic comparison in the sense that it manages to discriminate between alternative explanations for the recalcitrant evidence as outlined in equation (5). For example, the probability that h is true and all auxiliary hypotheses are false is identical with the alternative explanation supposition that h is false and all a_i's true.

The chosen response to these problems is to employ a Bayesian framework to compare the posterior probabilities of mutually exclusive explanations of the false predictions in the light of the same evidence (Howson and Urbach, 1989). Specifically, one explanation, that is, a conjunction of h and a_i consistent with $\neg e_i$, is preferred to an alternative provided its posterior probability exceeds that of the latter. Double application of Bayes' theorem implies that the relevant comparative posterior probabilities pertaining to any laboratory reproduction i are given by the following equations:

[10] Formally, if h and a_i are stochastically independent, then $p(h|a_i)=p(h)=p(h|\neg a_i)$.

$$(7)\ \forall i\ p(h \wedge \neg a_i | \neg e_i) \underset{<}{\overset{>}{=}} p(\neg h \wedge a_i | \neg e_i) \Leftrightarrow \frac{p(\neg e_i | h \wedge \neg a_i) p(h \wedge \neg a_i)}{p(\neg e_i)} \underset{<}{\overset{>}{=}} \frac{p(\neg e_i | \neg h \wedge a_i) p(\neg h \wedge a_i)}{p(\neg e_i)}$$

$$(8)\ \forall i\ p(h \wedge \neg a_i | \neg e_i) \underset{<}{\overset{>}{=}} p(\neg h \wedge \neg a_i | \neg e_i) \Leftrightarrow \frac{p(\neg e_i | h \wedge \neg a_i) p(h \wedge \neg a_i)}{p(\neg e_i)} \underset{<}{\overset{>}{=}} \frac{p(\neg e_i | \neg h \wedge \neg a_i) p(\neg h \wedge \neg a_i)}{p(\neg e_i)}$$

$$(9)\ \forall i\ p(\neg h \wedge a_i | \neg e_i) \underset{<}{\overset{>}{=}} p(\neg h \wedge \neg a_i | \neg e_i) \Leftrightarrow \frac{p(\neg e_i | \neg h \wedge a_i) p(\neg h \wedge a_i)}{p(\neg e_i)} \underset{<}{\overset{>}{=}} \frac{p(\neg e_i | \neg h \wedge \neg a_i) p(\neg h \wedge \neg a_i)}{p(\neg e_i)}$$

Consider, without loss of generality, equation (7). The first expression on the left hand side of the equivalence sign - $p(h \wedge \neg a_i | \neg e_i)$ - depicts the posterior probability that the main hypothesis h is true and the auxiliary hypothesis a_i is false, given the recalcitrant evidence $\neg e_i$. As can be seen from the first expression on the equivalence sign's right hand side, this posterior probability is a function of first, the likelihood of the disconfirmative evidence, given the conjunction ($h \wedge \neg a_i$); $p(\neg e_i | h \wedge \neg a_i)$ and, second, the prior probability of the outcome ($h \wedge \neg a_i$); $p(h \wedge \neg a_i)$[11]. The product of these two terms is then divided by the prior probability of the disconfirmative observation; $p(\neg e_i)$. Also, standard probability calculus implies that the probability of $\neg e_i$ can be expanded as follows:

$$(10)\ \forall i\ p(\neg e_i) = p(\neg e_i | h \wedge \neg a_i) p(h \wedge \neg a_i) + p(\neg e_i | \neg h \wedge a_i) p(\neg h \wedge a_i) + p(\neg e_i | \neg h \wedge \neg a_i) p(\neg h \wedge \neg a_i)$$

Consequently, the calculation of the comparative posterior probabilities listed in equations (7)-(9) hinges on the assignment of probability values to the three conditional likelihoods; $p(\neg e_i | h \wedge \neg a_i)$, $p(\neg e_i | \neg h \wedge a_i)$ and $p(\neg e_i | \neg h \wedge \neg a_i)$ and the alternative conjunctions of h and a_i[12].

A central part of this discussion is that Duhem's and Quine's theses come into play with regard to the likelihoods of the recalcitrant evidence. As stated above, Duhem's thesis basically asserts that logic alone cannot pinpoint the element(s) within a conjunction of a main hypothesis and auxiliary assumption(s) responsible for a false prediction. One interpretation of Duhem's position is to allocate equal prior probabilities to alternative likelihoods of the recalcitrant evidence. In contrast, Quine's thesis explicitly advocates granting central elements of the theoretical cluster immunity from recalcitrant evidence. Moreover, Quine posits a hierarchy of alternative ways of revision in the light of

[11] Prior probability does not mean an *a priori* point value, but, rather, a probability assigned prior to taking the observed evidence into account.

[12] The possibility of h and a_i being stochastically independent is disregarded.

conflict with experimental evidence in the sense that corrections are to be chosen with the intention of disturbing the total system as little as possible. One reading of this position is that the probability of the recalcitrant evidence if h is true and a_i is false supersedes the likelihood that a false h and a true leads to $\neg e_i$, which again ought to be greater than the likelihood that the false prediction is due to both a false h and a false a_i[13]. Table 2 contains a summary of assigned likelihoods of recalcitrant evidence that arguably reflect Duhem's and Quine's theses. Note that in the Quinean case α and β are positive constants greater than one, and $q \in (0,1)$.

***Table 2*. Duhemian and Quinean assignments of prior probabilities to likelihoods**

Likelihood	Duhemian priors	Quinean priors
$\forall i\ p(\neg e_i \mid h \wedge \neg a_i)$	q	$\alpha \cdot \beta \cdot q$
$\forall i\ p(\neg e_i \mid \neg h \wedge a_i)$	q	$\beta \cdot q$
$\forall i\ p(\neg e_i \mid \neg h \wedge \neg a_i)$	q	q

Furthermore, the prior probabilities of the conjunctions of main and auxiliary hypotheses are supposed to be equal. This is consistent with the Principle of Indifference, which entails assigning a probability of $1/n$ to each option in a situation characterized by n exclusive and exhaustive alternatives if there is no *a priori* reason to assign different probabilities to the various outcomes (also see Strevens, 2001). One practical interpretation of this approach is to consider each experimental reproduction a chance process that renders each conjunction of main and auxiliary hypotheses equally probable:

$$(11)\ \forall i\ p(h \wedge a_i) = p(h \wedge \neg a_i) = p(\neg h \wedge a_i) = p(\neg h \wedge \neg a_i) = 0.25$$

The priors outlined in Table 2 and depicted in equation (11) mean that the prior probabilities of the conjunctions of main and auxiliary hypotheses and the (implied positive) prior probability of recalcitrant evidence in equations (7)-(9) cancel out. Therefore, any comparison of posterior probabilities is completely determined by the likelihood ratios (Forster and Sober, 2001). Two conclusions immediately follow.

[13] Noticeably, the Quinean assignment is not at odds with the convention in the literature. When discussing a Bayesian approach to the Duhem-Quine thesis, Howson and Urbach (1989, pp. 99-100) employ nearly similar relative values of likelihoods (α=2, β=1, q=0.01).

First, Duhem's thesis is trivially incommensurable with unequal posterior probabilities. Equivalently, the experimental evidence does not favor any explanation of the false prediction over another. This is true for any reproduction of the experiment that produces recalcitrant evidence. Hence, a Duhemian assignment of likelihoods implies that conducting a series of reproductions does not diminish the problem of non-falsifiability.

Second, Quine's thesis implies that the recalcitrant evidence generally favors the explanation that the main hypothesis is true and the auxiliary hypothesis is false compared with the residual possibilities. Also, the explanation that the h is true and a_i is false will tend to have a higher posterior probability than the hypothesis that both h and a_i are at fault. Significantly, the solitary effect of conducting a series of experimental reproductions is to reinforce these claims. To see this, first note that the experimental reproductions are judged stochastically independent. This means that the conjoined posterior probability of any of the three explanations for a false prediction outlined in equation (5) is simply the product of the posterior probabilities pertaining to the individual reproductions. Consequently, a series of consistent recalcitrant evidence transforms the comparative posterior probabilities shown in equations (7)-(9) into strict inequalities consisting of multiple posterior probabilities as follows:

$$(7')\ \prod_{i=1}^{n} p(h \wedge \neg a_i | \neg e_i) > \prod_{i=1}^{n} p(\neg h \wedge a_i | \neg e_i) \Leftrightarrow (\alpha \cdot \beta \cdot q)^n > (\beta \cdot q)^n \Leftrightarrow \alpha^n > 1$$

$$(8')\ \prod_{i=1}^{n} p(h \wedge \neg a_i | \neg e_i) > \prod_{i=1}^{n} p(\neg h \wedge \neg a_i | \neg e_i) \Leftrightarrow (\alpha \cdot \beta \cdot q)^n > (q)^n \Leftrightarrow (\alpha \cdot \beta)^n > 1$$

$$(9')\ \prod_{i=1}^{n} p(\neg h \wedge a_i | \neg e_i) > \prod_{i=1}^{n} p(\neg h \wedge \neg a_i | \neg e_i) \Leftrightarrow (\beta \cdot q)^n > (q)^n \Leftrightarrow \beta^n > 1$$

The transformed equations reveal that each inequality only increases exponentially as the number of experiments grows larger. Therefore, a Quinean assignment of likelihoods means that a series of experimental reproductions of recalcitrant evidence affects just the degree to which retaining the main hypothesis and refuting the auxiliary hypotheses is preferred to the residual possibilities[14].

[14] This is a general result although equations (7')-(9') only compare the alternative consistent explanations for the recalcitrant evidence. A series of n experimental reproductions, each of which produces recalcitrant evidence, is consistent with 3^n feasible combinations of the possible conjunctions ($h \wedge \neg a_i$), ($\neg h \wedge a_i$) and ($\neg h \wedge \neg a_i$). Each series of conjunctions constitutes a possible explanation for the recalcitrant evidence. However, the assumptions listed in Table 2 and in equation (11) imply that in the Quinean setting none of the alternative series of conjunctions can attain a greater posterior probability than the consistent explanation that asserts that in the case of each experimental reproduction h is true and a_i at fault.

This argument is somewhat at odds with Smith's position in that it appears to limit the utility of reproducing experiments. Nevertheless, it should be noted that the Bayesian position developed above is clearly sensitive to the assignment of prior probabilities of the conjunctions of the main and auxiliary hypotheses. Here any experiment is deemed a chance process in the sense that the probabilities of all constellations of h and a_i are supposed to be equal and constant over the course of the laboratory reproductions. This approach may prove too restrictive relative to the prevalent practice in experimental economics alluded to by Smith. By way of example, reflect on the auxiliary hypothesis of adequate financial incentives (subject pool). In practice, experimentalists may be inclined to (act as if they) assign a higher prior probability to a main hypothesis tested in an experiment characterized by substantive financial incentives (economics students), as compared to a similar test employing miniscule potential financial rewards (history students) (confer Binmore, 1999 for arguments of this type). Conversely, assigning a low prior probability to a conjunction of a true main and a true auxiliary hypothesis is consistent with Starmer's (1999) argument that some experiments may be poorly controlled and/or do not constitute a relevant domain of the relevant theory. Moreover, experimentalists may (act as if they) revise prior probabilities during the course of running a series of experiments, as the stream of data updates the state of experimental knowledge (for an early and explicitly Bayesian analysis of experimental data, see Smith, 1964). A final point is that experimentalists may subscribe to likelihoods that differ from the values listed above in table 2 (Hylland, 2002).

Taken together, the employment of different subjective degrees of beliefs in the alternative (conjunctions of) hypotheses and likelihoods is arguably equivalent to Duhem's point that "good sense" is necessary in dealing with the recalcitrant evidence and non-falsifiability. By the same token, the ability to learn from the experience provided by experimental reproductions seems to require the specification of degrees of confidence in various hypotheses (also see Hausman, 1992). However, a formidable problem is that no intersubjectively agreed-on procedure for determining such prior probabilities exists (Worrall, 1994). By implication, no interpretation of experimental reproductions grounded by a set of subjective priors may then claim to be conclusive. In this sense the Duhem-Quinean problem of non-falsifiability persists.

Moreover, the discussed procedure of reproducing experiments in order to locate the source(s) of falsity is essentially a falsification procedure. Thus, another line of criticism might be to question the extent to which falsification is widespread in experimental economics. To be sure, when discussing the relevance of the Duhem-Quine thesis, Vernon Smith noted that

"Experimentalists and other economists often use the rhetoric of "falsifying" theories, but it is clear from the totality of our professional conduct that falsification is just a means to a different end: modification in the light of evidence" (Smith 1994, p. 129).

One reading of Smith is that the function of recalcitrant evidence extends beyond signaling falsity of main and/or auxiliary hypotheses. Specifically, anomalous series of laboratory data may assume a constructive and generative role in that they spur modifications of theory which aim at accommodating the false prediction. One example is the propounded hypothesis of individual risk aversion, instead of general risk neutrality, to organize data from first-price sealed bid auction experiments (for this and other examples of a similar kind, see Smith, 1989). Formally, a new hypothesis h' can be developed, which conjoined with the original auxiliary hypothesis a implies the false prediction:

$$(13)\; h' \wedge a \rightarrow \neg e$$

Even so, the positions of both Duhem and Quine merely state that the false prediction $\neg e$ can be inactivated by refuting h, a or indeed both h and a. Explaining the recalcitrant result is another matter that arguably extends beyond the non-falsifiability problem identified by the Duhem-Quine thesis (Quine, 1992, p. 16).

5. Conclusion

The Duhem-Quine thesis asserts that scientific hypotheses are always propounded for testing in logically interconnected clusters. Recalcitrant evidence entails the falsification of one or more hypotheses inside these networks, but logic alone cannot locate the particular element in a theoretical cluster responsible for a false prediction. These assertions are usually referred to as non-separability and non-falsifiability, respectively.

The purpose of this paper has been to reinterpret the relevance of the Duhem-Quine thesis for experimental economics. The first result is that non-separability applies to solitary laboratory experiments in economics. Theory-driven (data-driven) experiments require auxiliary assumptions that specify the environment, institution and residual features of the experimental design in order to derive an empirically observable prediction (benchmark) against which laboratory data can be compared.

Another issue is whether experimental methods can be used to diminish non-falsifiability by reproducing experiments and thereby stress-test a variety of auxiliary hypotheses. A Bayesian framework is employed to argue that reproductions do not solve the problem of non-falsifiability per

se. In particular, a close reading of Duhem's thesis appears to render any number of reproductions indecisive, whereas a detailed interpretation of Quine's thesis generally advocates retaining a main hypothesis at the expense of refuting auxiliary assumptions. Running additional experiments merely reinforces this Quinean recommendation. However, assigning different subjective degrees of beliefs to various conjunctions of main and auxiliary hypotheses may alter these conclusions. Then again, introducing subjectivity in this manner may serve to decrease the prospect for any unambiguous explanation for the recalcitrant data, whence the Duhem-Quine thesis perseveres.

References

Ariew, R. (1984): The Duhem Thesis, *British Journal for the Philosophy of Science* **35**, 313-325.

Binmore, B. (1999): Why experiment in economics? *The Economic Journal* **109**, F16-F24.

Cartwright, N. (1991): Replicability, Reproducibility, and Robustness: Comments on Harry Collins, *History of Political Economy* **23**, 143-155.

Cross, R. (1982): The Duhem-Quine Thesis, Lakatos and the Appraisal of Theories in Macroeconomics, *Economic Journal* **92**, 320-340.

Cross, R. (1998): "The Duhem-Quine Thesis" in Davis, Wade Hands, and Mäki (eds.): *Handbook of Economic Methodology*, Edward Elgar.

Davis, D. and C.A. Holt (1993): *Experimental Economics*, Princeton University Press.

DeWald, W.G., J.G. Thursby, and R.G. Anderson (1986): Replication in Empirical Economics, *American Economic Review* **76** (4), 587-603.

Duhem, P. (1991): *The Aim and Structure of Physical Theory*, Princeton University Press. (Reprinted version of the 1954 edition, translated by P. Wiener.)

Forster, M. and E. Sober (2001): "Why Likelihood?" forthcoming in M. Taper and S. Lele (eds.): Likelihood and Evidence, University of Chicago Press.

Gilles, D. (1993): *Philosophy of Science in the Twentieth Century. Four Central Themes*, Blackwell.

Grünbaum, A. (1976): "The Duhemian Argument" in Harding (ed.): *Can Theories be Refuted? Essays on the Duhem-Quine Thesis*, Reidel.

Hausman, D. (1992): *The inexact and separate science of economics*, Cambridge University Press.

Heijdra, B.J. and A.D. Lowenberg (1986): Duhem-Quine, Lakatos and Research Programmes in Economics, *The Journal of Interdisciplinary Economics* **1**, 175-87.

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

Hylland, A. (2002): Duhem, Quine og Bayes, Manuscript, Department of Economics, University of Oslo.

Jeffrey, R. (1989): *Formal Logic. Its Scope and Limits*, McGraw-Hill.

Lakatos, I. (1978): *The Methodology of Scientific Research Programmes*, Cambridge University Press.

Laudan, L. (1976): "Grünbaum on 'The Duhemian Argument'" in Harding (ed.): *Can Theories be Refuted? Essays on the Duhem-Quine Thesis*, Reidel.

Nickles, T. (1988): "Reconstructing Science: Discovery and Experiment" in Batens and van Bendegem (eds.): *Theory and Experiment*, Reidel.

Quine, W.V.O. (1980): *From a Logical Point of View*, Harvard University Press.

Quine, W.V.O. (1992): Pursuit of Truth. Revised version, Harvard University Press.

Roth, A. (1994): Let's Keep the Con out of Experimental Econ: A Methodological Note, *Empirical Economics* **19**, 279-289.

Sawyer, K. R., C. Beed, and H. Sankey (1997): Underdetermination in economics. The Duhem-Quine thesis, *Philosophy and Economics* **13** (1), 1-23.

Smith, V.L. (1964): Effect of Market Organization on Competitive Equilibrium, *Quarterly Journal of Economics* **78** (2), 181-201.

Smith, V.L. (1982): Microeconomic Systems as an Experimental Science, *American Economic Revie*w **72** (5), 923-55.

Smith, V.L. (1989): Theory, Experiment and Economics, *Journal of Economic Perspectives* **3** (1), 151-169.

Smith, V.L. (1994): Economics in the Laboratory, *Journal of Economic Perspectives* **8** (1), 113-131.

Starmer, C. (1999): Experiments in economics: should we trust the dismal scientists in white coats? *Journal of Economic Methodology* **6** (1), 1-30.

Steinle, F. (1997): Entering New Fields: Exploratory Uses of Experimentation, *Philosophy of Science* **64**, S65-S74.

Strevens, M. (2001): The Bayesian Treatment of Auxiliary Hypotheses, *British Journal for the Philosophy of Science* **52**, 515-537.

Vuillemin, J. (1986): "On Duhem's and Quine's theses" in I.E. Hahn and P.A. Schilpp (eds.): The Philosophy of W.V. Quine, La Salle, Illinois: Open Court.

Worrall, J. (1994): "Falsification, rationality, and the Duhem problem: Grünbaum versus Bayes" in J. Earman, A.I. Janis, and G.J. Massey (eds.): Philosophical Problems of the Internal and External Worlds: Essays on the Philosophy of Adolf Grünbaum, Pittsburgh: University of Pittsburgh Press.

Zahar, E. (1983): The Popper-Lakatos Controversy in the Light of 'Die beiden Grundprobleme der Erkenntnistheorie', *British Journal for the Philosophy of Science* **34**, 149-171.