Here is a prospectus for our (w. Harold Kincaid) session.

Session 1: Models, Evidence and Progress in Economics concerns the methodology of evidential reasoning and modelling in economics. Harold Kincaid discusses one of the central problems in philosophy of science, the Duhem-Quine problem. His main contributions include (a) a clarification of what the problem really is, which is often misunderstood in the current philosophy of economics literature, and (b) an illumination of scientific practices in economics drawing on (a). Kincaid argues that one cannot infer from the fact that hypotheses cannot be tested without auxiliary hypotheses to the claim that all scientific testing is holistic. Given that the Duhem-Quine problem arises in different ways in different contexts, one has to consider the contextual details to properly discuss the matter.

Economists are criticized for relying on highly unrealistic assumptions and the legitimacy of those assumptions has been one of the central themes in the philosophy of science. Aki Lehtinen argues that generality in economic models in theoretical economics bears epistemic benefits akin to what robustness analysis provides. That is, generality, like robustness analysis, alleviates problems arising from the use of unrealistic assumptions in economics. Generalization may provide epistemic advantages either by showing that some result from economic models are independent of the details of auxiliaries. Lehtinen also argues that generalization may help in allocating confirmation on particular assumptions in models. He (Lehtinen 2016; forthcoming) has already proposed a framework that explains how Duhem-Quine problems are solved when they are solved: the robustness of model-results allows for determining which results depend on which assumptions. In this paper, he shows that generality can be used for similar purposes, but that generality is better than robustness in that the falsities of the auxiliaries are removed rather than shown to be irrelevant. The papers engage with economic practices while drawing on substantial philosophical resources that have been developed in the post-Kuhnian philosophy of science. Kincaid considers empirical work in behavioural economics and Lehtinen discusses the use of the Dixit-Stiglitz model in international trade as examples.

Lessons of the Duhem-Quine Problem for Economics

Harold Kincaid

There is a well-known puzzle about scientific inference called the Duhem-Quine (D-T) problem which is source of continuing debate by philosophers of science and by some economic methodologists and economists. The problem is that hypotheses are usually (always?) tested with the aid of other auxiliary assumptions, making it hard to know where to assign blame or credit when the evidence is in. This paper surveys the state of play in thinking about the D-T problem and its significance for understanding testing in economics.

The paper is divided into two parts. Part 1 clarifies issues and surveys various responses to the D-T problem. The issues are complicated, and discussions in the context of the problem in economics have not taken advantage of the full range of proposed solutions in the literature. Part 2 address the problem as it surfaces in economics in light of the discussion in Part 1.

Part 1 first sorts out exactly what the D-T problem is supposed to show, pointing to an unfortunate tendency to run together two different theses, viz. 1.) that testing always involves other assumptions beyond the hypothesis at issue and 2.) that only whole bodies of assumptions are tested, never individual hypotheses. 2.) does not follow from 1.) as I illustrate in working through proposed solutions. I consider Bayesian (Dorling 1979) and likelihood (Sober) attempts to avoid acquiescing in the acceptance of 2.) with its nihilistic epistemic consequences. My general approach is what I would call contextualism.

Part 2 then applies the morals of Part 1 to the D-T problem in economics. I apply the results of Part 1 to some ongoing empirical work in economics, looking in detail at how the various proposed solutions to the D-T problem illuminate practices in economics, focusing especially on experimental work on the theory of choice.

This second section outlines the Duhem-Quine problem and solutions to it.Two standard proposed solutions to the Duhem-Quine problem come from Bayesian and likelihood approaches to confirmation. So Dorling (J. 1979, "Bayesian Personalism, the Methodology of Research Programs, and the Duhem-Quine Problem, " Studies in the History and Philosophy of Science 10: 177-187) followed by Howson and Urbach (1993, Scientific Reasoning Open Court, Peru Indiana) provide Bayesian "solutions" that are fairly obvious. The key is to be able to estimate separately the components of Bayes' theorem for H and for A. So we need to know P(H/e) versus P(e) and P (A/e) versus P(e). P(e) = P(E/H) P(H) = PE/~H) P(~H) and P(E/H) =P (E/H and A) p(A) + P (E/H and ~ A) p(~A). This is assuming that H and E are independent, i.e. P(A/H)= P(A). If we have these probabilities, then it is entirely possible that the evidence E (or ~E) tells differently against H and against A. If there is much evidence for H, little for A, and no obvious alternative to explain ~E, for example, then blame is clearly differential.

While formally correct, the Bayesian solution is subject to all the standard objections to subjective Bayesianism, with the the most basic being where do the probabilities come from (Mayo 2006) and what kind of epistemic normative force do conclusions based on subjective probabilities have (Sober 2004)? Of course this is a long debate with too much rhetoric and vitriol on both sides. Still, for economic applications, it is pretty easy to see that assigning prior probabilities to expected utility hypotheses and to the many assumptions used to test them (correct functional forms, etc.) is not at all obvious, not to mention assigning probabilities to the catchalls ~H and ~A. The normative force issue likewise surfaces when you consider the extant disciplinary trends and fads that there is good reason to think are misguided: whose probabilities are ruling here?

A second approach to the Duhem-Quine problem comes from Sober's likelihood inference

approach (2004, Journal of Philosophy). It tries to avoid subjective Bayesian problems by sticking to just objective likelihoods (no priors) and to specific competing hypotheses to avoid problematic catchalls. The end result is relative comparisons of hypotheses' plausibility, with no claim to some measure of absolute believability like the Bayesians hope for. Applied to the Duhem-Quine problem, the likelihood approach can distinguish support of two specific hypotheses.

The problem with Sober's solution--which he openly admits--is that it requires we have independent objective data about the probability of seeing the observed data for each combination of hypothesis and auxiliary and that we know what hypotheses and auxiliaries to compare, where they are not Bayesian catchalls of the form ~H and ~A. These are strong demands that will often be hard to meet with economic data.

What are we to make of these two more or less formal alleged solutions to the Duhem-Quine problem? Granting the objections to the Bayesian solution, it does make two compelling points. Foremost, the problem dissipates if we have evidence about H and about A that are independent of each other. Second, assessing evidential blame or credit calls for looking at the evidence for alternatives to H and A. We will use these points when we turn to look at evidence for choice theory.

This paper adopts the stance that there is not going to be one simple solution to the Duhem-Quine problem. Thus I turn in the second section to apply these morals to the Duhem-Quine problem in economics.

Part II Duhem Quine and Economics

Multiple authors have said that the Duhem-Quine problem is significant for science in general and in particular for economics. See for example Hands 2001 and Bardsley et. al 2011.

The idea that hypotheses in economics are only tested with the help of auxiliaries is slogan that hides a great deal of complexities in the As. H & A entails E is logical formalism that does not get us

very far. The paper sorts out in some detail a number of different kinds of auxiliaries that are unlikely to have the same status and to be assessed in the same way:

Statistical assumptions

Model assumptions

Causal assumptions

Competitor assumptions

Idealizing assumptions

It is also important to realize that it is often case that there are multiple hypotheses being tested even excluding the rough categorization of auxiliary or testing assumptions.

The general strategy for testing such assumptions is that outlined above: look for evidence that does not depend on the the truth of the hypothesis (hypotheses) at issue. There are myriad ways of doing so. E.g.

Show that the auxiliaries assumed are not essential to testing the hypothesis at issue--that alternative auxiliaries would allow the same implications about observed evidence.

Show that predictions get more accurate as auxiliaries are made more realistic

More routes will be described in the paper produced after the conference and mentioned in the presentation.

Evaluating work in economics in terms of its ability to test auxiliaries successfully involves two key components: what does the theory show and what are the practices of researchers? The theory may be neutral on auxiliaries while researchers themselves systematically fail to use available methods to

test them.

These general points are then used to look at practices and developments in experimental and observational studies of choice theory. I first look at what seems to be best practices and their evolution over the last decade or so. With that in hand I then look at the "pseudoscientist" side of experimental and observational work on the the theory of choice, i.e. at standard practices that are more or less guaranteed to not rigorously test hypotheses and their associated auxiliaries.

Best practices for attributing credit and blame include providing evidence for these auxiliaries, among many others:

doing time only jointing estimating with risk attitudes
incentive compatible designs
testing magnitude effects
front end delays
mixture models
testing stability
order effects
mixed lotteries over loss and gain frames
testing any given theory of choice against most serious competitors
making clear what part or parts of any given choice theory any particular piece of evidence is relevant to
This list can certainly be expanded.

I list a number of practices where it is theoretically possible to provide differential evidence for hypotheses against auxiliaries, but common practice skirts those opportunities.

Turning next to the substandard practices in experimental work on choice theory, there are numerous strands to point to:

- Tests of choice theories using hypothetical questions, not real payoffs. There is lots of evidence that using the auxiliary assumption that hypothetical questions get at true preferences is mistaken, but the practice continues.
- Tests of discount rates that do not take into consideration risk attitudes. This is fairly common, but there good theoretical reasons to think that risk attitudes are an essential auxiliary for most tests of discount rates.
- Hypotheses are tested in race horse fashion, invoking the auxiliary assumption that one model must describe the entire population. That auxiliary assumption can be wrong and there are wasys of testing it, but they are gerally not used.

Many more such willful failures to test such auxiliary assumptions can be identified.

So the upshot is that while there is no all purpose answer to the Duhem Quine problem, it can be answered case by case, depending on what background knowledge we have and what methods we have at our disposal. That is true in general, and it is true in economics.

TITLE: The epistemic benefits of generalization in economic modelling

ABSTRACT: This paper spells out why generality is an important desideratum in economic modelling. Generalising models serves similar epistemic functions as robustness analysis: it provides a solution to the epistemic uncertainty that arises from the presence of unrealistic assumptions. We present our arguments by discussing examples from economic modelling: the Dixit-Stiglitz (1977) model of monopolistic competition and Abraham Wald's proofs for the existence of general equilibria.

Modern theoretical economics largely consists of building, manipulating and modifying abstract and idealised mathematical models. Proving the same result with more general assumptions is an important achievement for economists. Despite the importance of generality in model-based economics, however, thus far there has been little reflection on what it is and what it is good for. In this paper, we examine generality as a modelling desideratum and the ways in which it is achieved, i.e. generalization.

Generality is typically defined as 'the property of applying widely' and is measured either by the number of phenomena a model can explain or predict, or by the number of systems to which a model applies (Lewis and Belanger 2015; Levins 1966; Orzack and Sober 1993; Weisberg 2004). But why is 'applying widely' a desideratum?

We will argue that not all kinds of generalizations provide epistemic benefits. Specifically, generalizations have epistemic benefits only when they involve either increases in expressive power or they entail fewer false assumptions about the target. Increasing a model's generality via either of these routes helps solving some of the problems that arise from the necessity of making unrealistic assumptions. Tractability considerations often imply that modellers describe their targets with assumptions that are less general than they think can be truly asserted about them. Thus they often are able to prove a result only for a special case. They know the specificities introduce falsehoods. Yet they do not know whether the model's results crucially depend on those falsities. When the model is generalized but continues to imply the same result, we learn that the particular falsehoods were not responsible for the results. That is, obtaining the same result with less restrictive assumptions increases the modellers' confidence that the result is not an artefact of specific assumptions that are known to be unrealistic. We will thus argue that the importance of generality derives from the same kind of epistemic considerations that motivate derivational robustness analysis (see esp. Kuorikoski, Lehtinen, and Marchionni 2010; Weisberg 2006; Wimsatt 1981), and that herein lies its main epistemic advantage.

There are three kinds of generality and corresponding generalizations. First, a model may be generalized such that it applies to more phenomena. Second, the level of abstraction of the target may be increased. Third, a given target may be described with more general assumptions. If we simply count the number of systems, the three cases are indistinguishable – in all the number of systems becomes larger. Let us call them *increasing the number of target phenomena* (G1), *generalizing the target* (G2), and *increasing the number of subsumed systems* (G3), respectively. We argue that the three kinds of generalizations are very different, and that only G3 provides genuine epistemic benefits. Intuitively, epistemically beneficial generalizations occur when the model-descriptions are modified in such a way that they capture a larger

In practice, model M_1 is more general than M_2 if it makes fewer assumptions than M_2 or if some of its assumptions have more expressive power. We may then present a characterization of epistemically beneficial generalizations:

> *Model M_1 provides an epistemically beneficial generalization of model M_2 if the model-implications of M_1 and M_2 include the same generalized (or actual) targets and M_1 describes them in such a way that they subsume a larger number of possible systems than M_2.*

A large part of the paper is devoted to making the necessary conceptual distinctions that are needed for expressing what this means exactly. Let $S=\{p_i,\ldots\}$ mean that system S has property p_i. There will be systems like this:

$S_1=\{p_1,p_2,p_3,\ldots,p_n\}$

$S_2=\{p_1,p_3,p_4,\ldots,p_n\}$

$S_3=\{p_1,p_4,p_5,\ldots,p_n\}$

$S_4=\{p_1,p_2,p_3,p_4,p_5,\ldots,p_n\}$

We will say that a target *subsumes* one or more systems, such that the cardinality of the subsumed systems is given by the number of systems that share the properties that define a *target*. A target thus picks out only selected properties from systems, those that the modeller intends to account for. For example, if the modeller only wants to account for p_1, such that T={p_1}, T subsumes all of the systems S_1 to S_4. Let us denote this relation as follows: T(S_1,S_2,S_3,S_4). If the target T_1 is defined by T_1={p_1,p_2,p_3}, then T_1(S_1,S_4). For T_2={p_1,p_4}, T_2(S_2,S_3,S_4), and so on.

A model M *applies* to system X if its model descriptions successfully represent the features that define the target T, and system X indeed has the properties that define the target. Thus, for example, model M_1 applies to systems S_1 and S_4 if its target is T_1={p_1,p_2,p_3}, and the model successfully describes properties p_1, p_2, and p_3. If it does successfully describe those properties, let us write $M_1 \vdash p_1, p_2, p_3$. We can now place the modeller's intentions on the left side of the symbol $\vdash$. The expression

$M_1(T_1) \vdash p_1, p_2, p_3$

means that the modeller intends M_1 to apply to target T_1, and it does apply to this target in virtue of successfully representing the properties that define the target. We can also write

$M_1(T_1(S_1,S_4)) \vdash p_1, p_2, p_3$

to indicate that the modeller intends to capture systems S_1 and S_4 by way of defining the target as T_1. Finally, we can write $M_1(T_1(S_1,S_4)) \vdash T_1$ and $M_1(T_1(S_1,S_4)) \vdash S_1$, S_4 or $T_1(S_1,S_4)$ to indicate that the model successfully applies to T_1 and thus applies to systems S_1 and S_4.

This definition of applicability is rather lax in terms of not specifying anything about how well a model applies to a target. Whatever fidelity criteria modellers use for deciding which features the target has, if the model descriptions are able to represent those features, it applies to a system that has those features. Importantly, this notion of applicability does not rule out model descriptions that are too specific. The model descriptions may mis-describe the systems subsumed by the target or even give a description that only truthfully applies to a subset of all systems that have the characteristics that define the target. Thus, for example, A model M may successfully apply to system S_4= {p_1,p_2,p_3,p_4,p_5,…,p_n} if the target is T_1={p_1,p_2,p_3} if

$M(T_1) \vdash p_1,p_2,p_3,$

But it also applies to S_4 if

$M(T_1) \vdash p_1,p_2,p_3,p_4,p_5$ and if

$M(T_1) \vdash p_1,p_2,p_3,p'_4,p'_5$, where p'_4 is incompatible with p_4, and p'_5 is incompatible with p_5.

The point here is that there may be a mismatch between the set of systems to which a modeller wants her model to apply to (e.g., here S_1 and S_2), and the way in which the model descriptions characterize them. The problem is that the model-result concerns the systems as characterized by the model descriptions rather than the systems as defined by the more abstract target. We can write the model-results as follows:

$M \vdash R(p_1,p_2,p_3,p_4,p_5)$

Here the idea is that the model result R holds in a system (S_4) which has properties p_1,p_2,p_3,p_4,p_5.

Intuitively, one can provide an epistemically beneficial generalization if one can prove a given result for a given target such that the target subsumes a larger number of systems due to making fewer assumptions about the target or by having more expressive power.

For a concrete example, consider Krugman's generalization of his (1980) model. The initial model assumes that population size is the same in the two countries (A_4). Krugman notes that relaxing this assumption by letting the population sizes be arbitrary does not affect the main result: 'It can be shown that in that case, although the derivations

become more complicated, the basic Home Market Result (HMR) [that each country exports the goods in which it has a large domestic market] is unchanged'. If X denotes the rest of the assumptions in the model, this generalization removes assumption A_4 so that the model changes from (K)

$$(A_1, A_2, A_3, A_4, X) \vdash HMR \qquad (K)$$

to (K'):

$$(A_1, A_2, A_3, \quad X) \vdash HMR. \qquad (K')$$

If the result crucially depended on the special assumption of equal population sizes, we would have reasons to think that it only holds in circumstances that never hold in reality, namely that it is merely an artefact of tractability assumptions. By generalizing the result in this way, Krugman shows that it is more likely to hold in reality. The main mechanisms and the main result remain the same, but they are described with more general assumptions. The functioning of the mechanism and its characteristic results have been shown to be independent of some of the details included in the original description.