INVESTIGACIONES ECONÓMICAS. vol. XXX (2), 2006, 179-206

HOW TO DO EMPIRICAL ECONOMICS

FRANCIS KRAMARZ (Editor)
INSEE
JOSHUA D. ANGRIST
MIT
DAVID M. BLAU
University of North Carolina
ARMIN FALK
University of Bonn
JEAN-MARC ROBIN
Université de Paris 1
CHRISTOPHER R. TABER
Northwestern University

This article presents a discussion among leading economists on how to do empirical research in economics. The participants discuss their reasons for starting research projects, data base construction, the methods they use, the role of theory, and their views on the main alternative empirical approaches. The article ends with a discussion of a set of articles which exemplify best practice in empirical work.

Keywords: Empirical research, econometric methods.

(*JEL* B4, C5, C8, C9)

1. Introductory note by the Guest Editor

I was asked by INVESTIGACIONES ECONÓMICAS to organize a discussion between leading empirical economists on "How to do empirical economics". In fact, the questions they answered were all addressing a more personal issue: "how do you practise your empirics". Is it a matter of taste? Are we all doing economics or are we heading towards a more general social science? Clearly, there is a lot of heterogeneity in the conceptions within the profession and we have tried to reflect this

heterogeneity in our choice of interviewees. I assume that the journal's editors also selected someone who might stand in an intermediate position within the *debate* to organize it. Indeed, thinking about social science *versus* economics may be a useful starting point when reflecting on an empirical strategy.

Models abound in economics. Testing them is rather natural, at least for some. For instance, structural estimation is a *natural* research strategy when interested in the job search model, an endeavour for which Jean-Marc Robin just received the Frisch medal. But theory is much less clear when one goes further away from economics and what matters is providing clear *facts*, i.e. clear and robust causal relationships. The quality of data and the quality of identification have become essential elements in our capacity to produce scientific evidence. There, Josh Angrist has led the pack and pushed all of us in directions that were unanticipated 10 or 20 years ago. Another debate revolves around data and the role of experiments. It was therefore a pleasure to have Armin Falk with us. A recent article by List and Levitt (2005) constitutes a good complement to Armin's response in our discussion. In particular, List and Levitt discuss strengths and limits of experiments in more detail than is possible here, because our questions are, in a sense, more personal. David Blau and Chris Taber offer balanced perspectives, using broad methodological approaches, even though both are strong believers in economic models. All of our interviewees are excellent econometricians who use the most advanced techniques, or even advance the techniques if they feel this is necessary for their empirical goals. They are all role models for empirical economists, even though we might sometimes disagree with one element or one detail of their research strategy. But, when reading them, we learn from them, even when we might have adopted a slightly different route.

2. Starting a project

2.1. Why do you start an empirical project? Is it mostly because you want to evaluate a public policy; because you want to test an economic theory; because you want to estimate a parameter, an elasticity that has a central role in a model; because you want to answer an economic or a social question; because you want to understand the micro-behaviour of agents; because you want to understand the macro-behaviour of an economy; etc.

ANGRIST: I usually start a research project because I get interested in a causal relationship. I put causal questions at the top of my agenda because the answers to these questions can be used directly for predicting economic outcomes and for policy analysis. For example, this year I have been working on quantity-quality trade-offs, i.e., the causal link between sibship size and child welfare. Clearly, theory is a big motivator here, with important contributions by Becker providing the main theoretical context. On the other hand, development policy all over the world is predicated on the notion that big families are bad. We got the one-child policy in China and forced sterilization in India, largely because of casual empiricism linking large families and rapid population growth with bad outcomes. I don't think Becker has much to do with this. With or without the theory, it is worth finding out whether these policies are misguided.

Sometimes I get interested in a particular causal question after learning about an institutional feature that suggests a new way to answer an interesting question. For example, I first learned about the draft lottery from Orley Ashenfelter, in his graduate Labor class. Orley came into class one day and described how he had heard about epidemiologists who had compared the civilian mortality rates of men who had been at high and low risk of serving in the Army due to the draft lottery. Orley said, "somebody should do that for these guys' earnings." So I went from Orley's class to the library and got to work. A later example in this spirit is my Maimonides' Rule paper with Victor Lavy. Victor and I decided to write a paper about class size after we discovered the ratcheting Maimonides pattern in the relation between class size and enrollment.

Sometimes my motivation for a project or question comes from earlier work and a desire to complete the picture. For example, Alan Krueger and I used to talk about whether World War II veterans really earn more or whether this is just selection bias, as suggested by the Vietnam results. We dug around in the government documents section of Princeton library until we came up with an instrumental variable (IV) strategy from birthday-based conscription. Later, I felt like I ought to have something to say about voluntary military service, since the earlier work on conscription naturally raised the question of whether the effects of voluntary military service are also negative. Then I learned about the ASVAB misnorming (when US military entrance exams were incorrectly scored) and that seemed to provide the solution.

BLAU: I start empirical projects for a variety of reasons. I might be inspired by a really good paper to replicate or extend the approach in the paper. Sometimes I notice something in data that does not appear to have received much attention but seems interesting (e.g. the end of a very long trend of decline in the rate of self-employment in the US). Given my interest in economics of the family, I might read studies by demographers or sociologists on an issue to which I could imagine applying economic reasoning and analysis (e.g. the impact of the cost of child care on employment behavior of mothers of young children). Sometimes I read a lot of papers on a subject, and I am not satisfied with the approaches taken in the papers or convinced by their results, and I imagine that I could do a better job (e.g. the retirement-consumption puzzle). I do not usually begin a project with the goal of evaluating a specific policy, but I think about the policy implications of the research from the beginning.

FALK: It is a mixture of curiosity and the desire to explore a socially or economically relevant question. It is rewarding to discover something and to test one's intuitions and hypotheses. This holds even more so, if the analyzed question is politically relevant. In general I think there is no shortage of intriguing questions. What I often find difficult to decide is which project to pursue first, or which ones not to pursue at all.

ROBIN: Chance plays a big role in determining an empirical project. A paper you refereed, a chat with a colleague, an idea you had let aside while working on a previous project, questions in seminars or from referees, etc.

I started my Ph.D. thesis on equivalence scales. The data I was using were a French survey on household food consumption. I had the data by chance, thanks to some particular advisor I had had before. From equivalence scales I switched to infrequency of purchase models because of the particular survey design. There was very little economic theory there but a lot of statistical modelling to produce inference on household consumption from household purchases. Later, I would try to design a structural (S,s) model of purchase renewals to fill the theoretical gap.

Chance again: Richard Blundell and Costas Meghir were also working on equivalence scales at that time (the end of the 1980's). I met them and started to work with Costas on some multivariate statistical model

of infrequent purchase. With Richard, we worked on designing simple econometric procedures to estimate demand systems. Because family expenditure data always aggregate consumptions at a certain level (for example, there is no way of determining which part of energy consumption that is used for heating, cooking or whatever), we proposed the concept of latent separability. And so on and so forth.

Chance plays a big role in triggering a new empirical project. Then, if you are lucky, a simple initial idea will give you work for many years. At some point, either you get fed up with a particular topic, or you feel you have said all you wanted to say and you start something else. I do not remember feeling the writer's block. This is because one has always many more ideas of papers than one can pursue. When I got bored with consumption econometrics I became interested in labour economics and I found other people to work with and other subjects very naturally.

TABER: All the points in the question may motivate me. I typically start an empirical project when I have a question that I find interesting and that I have an insight on how to answer it. The questions might arise from three different sources. First, it could be something that I came upon by accident. I may be reading a paper or attending a seminar and it strikes me that things can be done better. Alternatively, I may come upon an idea more purposely–I start with a general area of interest and read the literature and see if I can improve on it. Third, a question might arise while working on a previous paper.

2.2. Now, about how you do empirical work. When you start a project, can you describe your general methodology?

ANGRIST: The two hardest things about empirical work are picking projects and knowing when to bail out on projects that are not developing well. When I am scouting a project for the first time, I read a lot, trying to find out what has been done. I worry that the question has been addressed before, and that even in the best scenario I'll have little to add. In the early stages, I also look for excuses to abandon a project, say a falsification test that will shoot it down. Another important hurdle is whether there is a plausible first-stage, broadly speaking. For example, Daron Acemoglu and I once set out to study the effects of advanced notice provisions (the requirement that workers be notified of impending layoffs). We could not find any evidence that laid-off workers actually got advanced notice, even though we had some nice

reduced forms for the outcome variables. So we put that one out of its misery. Of course, it is not always so clear-cut. Sometimes setbacks are temporary and I misjudge their severity. I often make mistakes and bail out too soon or too late.

BLAU: I read existing studies carefully and write a summary of findings, limitations, and what we would like to know. I often seek a grant to support a new research project, and I find that writing a coherent and convincing proposal helps me focus my ideas. I write down a simple theoretical model to help clarify the key issues. Derive hypotheses from the theory, if possible. Explore the implications of the theory for the data needed and the empirical approach. Look at what is available in the most obvious data sources. Revise the empirical approach in light of limitations in the data.

FALK: At the beginning there is always an idea. This idea can come from many sources, e.g., reading papers, attending seminars or discussing with researchers. Besides these rather *traditional* sources they can in principle come from anywhere. Since I am interested in the psychological foundation of economic behavior, all kinds of social interactions in my daily life shape my research agenda. In this sense data collection and idea generating is intimately related to almost everything I experience. Once the idea is born it is critically questioned in terms of novelty and relevance. I try to find out how exciting or intuitive the idea is. If it passes this test, I start designing an experiment to test the idea. Of course, what I just described as a well-defined sequence of events may well be coordinated but also chaotic, may happen in a second or over years and sometimes leads to something useful and sometimes nowhere.

ROBIN: There is not *one* methodology. In some cases, I only wanted to produce evidence. For example, I worked on infrequency of purchase models to produce evidence on consumption given purchase data. I worked on lifetime inequality because I wanted to produce a synthetic picture of both cross-section earnings inequality and earnings mobility. Other projects have deeper roots in theory. I worked on designing equilibrium search models with auction-type wage setting mechanisms because I thought that the Burdett-Mortensen model was making assumptions that were not quite right. And I am currently doing some theoretical econometric work on independent factor models because I think that the identification of microeconometric models with multi-

dimensional sources of heterogeneity is bound to become a very important topic in the near future.

So sometimes my work is very descriptive, other times it seems determined by theoretical considerations or it looks like statistical methodology. Now, my deep motivation is always for application. I have hardly written a single paper without an application on actual data.

TABER: Of course this varies across paper. As a general rule I try to start by first writing down an econometric model of the general problem (usually substantially simplified). Within the context of the model I think about the goal of the empirical project. I then think about the issue of the conditions under which that effect can actually be identified from the type of data I am likely to get.

2.3. Going into details.

a. More precisely, how long do you spend assembling and constructing the data sources?

ANGRIST: It is usually more satisfying to work with new data than to, say, run some new regressions with the National Longitudinal Survey of Youth (NLSY) or even the Current Population Survey (CPS). Of course, new data is also more work. But the odds of having something exciting to say go up considerably when the data are new, and go up even more when you have constructed the data set to serve your particular agenda as opposed to being limited by someone else's idea of what the world's research agenda needs. Also, with new data, the odds someone else will beat you to the punch go down. Another consideration is that for the type of work I do, most off-the-shelf data sets are too small. As far as IV strategies go, for example, I cannot imagine getting much of substantive value out of the Panel Study of Income Dynamics (PSID) or NLSY, though they are fine for econometricians to practice their chops on. Among public-use data sets, I especially like the Public Use Microdata Sample files, because of their size and simplicity. But it can still take a long time to put these together because of changes across years, or the need to link within families.

BLAU: It clearly depends on the data. For some projects involving complicated data (e.g. Health and Retirement Study matched to employer-provided pension and health insurance records, and administrative Social Security records), two years. For others, a few months (e.g. CPS, NLSY) are enough.

FALK: My preferred research tools are experiments, both in the lab and in the field. I value this method so highly because it offers a unique possibility to control for confounding factors and allows causal inferences. The control possibilities available in laboratory experiments go substantially beyond the respective controls in the field. In a well-designed experiment you control the strategy sets, the information sets, payoffs, technology, endowments, framing etc. In an experiment you know quite well which variables are exogenous or endogenous, you can implement exogenous treatment variations, you can study the degree and the dynamics of equilibrium adjustments and, if someone doesn't believe your results, he can easily replicate the experiment, establishing solid empirical knowledge.

Some critics of experiments, however, worry about so-called *external validity.* Lab experiments are often criticized to be unrealistic because of a potential subject pool bias (undergraduate students) or the relatively low stake levels used in experiments. Moreover, it has been pointed out that subjects typically know that they are acting in an experiment, and that their actions are observed by an experimenter, which may induce unrealistic behaviour. In addition, in most economic experiments subjects typically choose numbers or points instead of, e.g., quality or effort levels.

What can be answered to this criticism? First, to me it is anything but clear what external validity really is and why it is a criticism at all, given that subjects in experiments are real people who take real decisions for real stakes. Second, all depends on your research question. Just like economic models, experiments are unrealistic in the sense that they leave out many aspects of reality. However, the simplicity of a model or an experiment is often a virtue because it enhances our understanding of the interaction of the relevant variables. Moreover, often the purpose of an experiment is to test a theory or to understand the failure of a theory. Then the evidence is important for theory building and not for a direct understanding of *reality.* Third, the honest sceptic who challenges the external validity of an experiment has to argue that the experiment does not capture important conditions that prevail in reality. The appropriate response is then to try to implement the neglected conditions. Fourth, field experiments (I am not talking about natural experiments here) offer a neat way to combine a relatively high level of control with an ecologically valid decision environment. Let me give you an example: In a recent study I collaborated with a

charitable organization, which allowed us to study the nature of social preferences in a controlled, and yet natural environment. We simply varied whether donators received a gift together with a solicitation or not and found that the larger the included gift the higher the donation probability (Falk, 2004). Thus it is possible to perform an experiment observing behaviour of a non-student subject pool, where participants do not know they act in an experiment, where the size of the stakes is not predetermined by an experimenter and where behaviour involves *real* items and not the choice of abstract numbers. Fifth, it is also possible to combine empirical methods to overcome the external validity critique and to make use of valuable complementarities. A good example is the combination of experiments and representative surveys. In a recent study on individual risk attitudes (Dohmen *et al.,* 2005) we analyze the responses of about 22,000 people of the 2004 wave of the German Socio Economic Panel (SOEP). Since survey questions are not incentive compatible they might not predict behaviour well. We therefore conducted a field experiment with 450 representatively selected subjects that answered the same questions and took part in a lottery experiment with real money at stake. It turns out that the responses to the questions reliably predict the behaviour in the lottery, which validates the behavioural relevance of the SOEP survey measure.

ROBIN: Not so much as I mostly used data assembled and constructed by other researchers.

TABER: This varies tremendously across projects on which I have been involved. Typically getting the data in the form that I need it to estimate the model at hand takes a long time. Of course, if it is a data set I have used before on a related problem, it will be much quicker.

b. Are these sources generic (e.g. the CPS, the PSID, etc.), produced by others, or designed, collected, and constructed by you, with the help of your research assistants? Can you give examples?

BLAU: Usually generic. In some cases with restricted access data merged in, which requires jumping through a lot of bureaucratic hoops to gain access to the data. I, and my research assistants, typically spend a lot of time extracting, examining, cleaning, and transforming data into a usable form. Particularly with longitudinal data, a lot of consistency checking is necessary. For example, I am using the NLSY to create a *co-residence* history incorporating both cohabitations and marriages. About 10% of the sample has apparently inconsistent histories (e.g.

report the end of a marriage but never reported getting married). An experienced programmer is going through these cases to look for coding errors and develop algorithms to correct the cases that can be fixed.

ROBIN: The French Labour Force survey, the French administrative source of data on workers' wages (DADS) or firm accounting data (BRN), the British Family Expenditure Survey, the French one.

TABER: I typically use generic data sets including the NLSY, CPS, National Education Longitudinal Survey of 1988 and Survey of Income and Program Participation. Often these data need to be augmented in some way with additional sources.

c. Do you start from a theoretical model, an econometric or a statistical model?

BLAU: A theoretical model. This is a necessary first step for me, to help focus my thinking and clarify the economic issues. I derive testable hypotheses from the theory, if possible. If I plan to take a structural approach to estimation, then I expand the theory to incorporate important institutional features.

FALK: My experiments are often designed to test theories. This implies that experiments are intimately related to the development of game theory. It is therefore no surprise that among non-experimentalists game theorists were the first to show much interest in experimental results. Some of the most exciting recent developments in applied microeconomics are inspired by laboratory findings. Therefore it is almost impossible to run good experiments without doing theoretical work as well.

TABER: I would say that I typically start with an econometric model. However, in some sense by definition an econometric model is also a theoretical model and a statistical model.

d. More generally, what is the role of economic theory in your favoured approach?

ROBIN: In general economic theory plays a huge role. I do not believe that a formal model is putting restraints on one's intuitive capability of thinking economic facts. Quite the contrary: a formal economic model not only helps to understand economic mechanisms better, it also helps to understand where individual heterogeneity should play the major part, the potential sources of simultaneity or selectivity biases. I know

that some people have a much better intuition of all this and need less formal economic modelling than me. I envy them.

TABER: This varies widely across what I am doing. Some of my work with Heckman and Lochner specifically examines equilibrium effects of the labour market. Economic theory is completely central to this work. Other work, such as my *Review of Economic Studies* paper looks at the returns to schooling so I always had models of Roy, Becker, and Mincer in the back of my head as I formulate the problem. However, on a day-to-day basis, economic theory did not play a central role. I have also worked on pure *treatment effects* papers in which economics plays essentially no role such as my work on Catholic Schools with Altonji and Elder. I can write down a human capital model to think about the problem, but I am not sure it really adds that much to the interpretation. I should say that although I don't think economic theory is important for all work in empirical microeconomics, I personally enjoy working on papers in which economic theory takes a large role.

e. What is the role of econometrics in your favoured approach?

ANGRIST: I like clever new econometric ideas as much as the next guy, maybe more. But econometrics for its own sake should not be confused with what I call *real empirical work*, which is question-driven. Most causal questions are better addressed using regression or Two-Stage Least Squares (2SLS) than fancier methods. This is because the case for causality is always so hard to make. Use of simple tools focuses your attention on core identification and measurement problems instead of second-order considerations like how to handle limited dependent variables. It also helps you avoid mistakes (though a number of famous papers get 2SLS wrong).

On the other hand, sometimes new econometric methods lead to a valuable simplification. An example of this is quantile regression for the analysis of effects on distributions. I prefer the quantile regression framework to kernel density methods or a direct effort to estimate distribution functions because all my old ideas about how regression works carry over to quantile regression in a reasonably straightforward way. It is also easy to get the standard errors.

BLAU: I use the appropriate econometric method for the problem at hand. This could be a simple linear model derived as an approximation to a decision rule implied by the theory, estimated by Ordinary Least Squares (OLS) or 2SLS (e.g. the effect of child care subsidies

on employment of mothers, using cross section data with county-level variation in subsidies), or Fixed Effects (e.g. the effect of income on child development, using longitudinal data). If the problem is more complicated, then I write down the likelihood function and figure out how to identify the parameters and maximize the function, still in the framework of *approximate decision rules* (e.g. a joint model of choice of a mother's decision to work and type of child care to use, with common unobservables in the two models). If I plan to estimate the model structurally, then I derive the likelihood function or other objective function from the model together with assumptions about distributions and functional forms.

FALK: In experiments econometrics are typically less important compared to using field data simply because, in a sense, the econometrics is built into the design. If I am just interested in simple treatment effects I prefer using simple non-parametric tests, which are best suited for the analysis of experimental data. If the interest goes beyond that I use standard econometric techniques, e.g., to simultaneously control for multiple factors or to study interaction effects.

ROBIN: I consider myself as an econometrician. I try to keep up with the latest techniques.

TABER: Econometrics has played a very large role on almost every empirical project on which I have ever worked.

f. Should the methods be simple or up-to-date?

BLAU: The methods should be appropriate for addressing the question of interest. A narrowly focused question in which generalizability and extrapolation out of sample are not of primary interest would typically call for a simple method that requires few assumptions. An issue for which more general results and out of sample extrapolation are of interest will usually need a more structured approach and a more sophisticated econometric method. Two examples from my research: 1) I wanted to know whether existing child-care regulations in the US increase the cost and quality of child-care. This is a relatively narrowly focused question, and I was not particularly interested in using the results to predict the effects of new regulations. I used simple linear difference-in-difference (across states and over time) methods. 2) I wanted to know whether lack of retiree health insurance affected the timing of retirement. There was an important policy issue involving Medicare, which provides public health insurance for the elderly in the

US An issue of interest was whether changing the age of eligibility for Medicare would affect the impact of retiree health insurance on the timing of retirement. But the age of eligibility for Medicare has been unchanged since the program began. A dynamic structural approach was needed in order to identify and estimate the degree of aversion to medical expenditure risk, and the implied impact of changing Medicare policy.

ROBIN: Adapted.

TABER: I suppose that I think they should be up to date in the sense that I think researchers should be aware of the current status of econometrics and use the methodology that is best for the problem at hand. However, all else equal, simple is obviously better. There is no reason to add econometric complications needlessly. However, I think quite often the appropriate methodology is not simple (or in some cases the best methodology is simple, but understanding why it is appropriate may be quite difficult).

3. Alternative empirical approaches

3.1. In some areas of research the descriptive approach is widely used (e.g. wage inequality and mobility, adjustments for quality in inflation measurement, job creation and job destruction statistics). What is the usefulness of this approach?

ANGRIST: Mostly, to generate new questions. Sometimes to provide context for answers. But I admit that a lot of purely descriptive work bores me – especially work that I cannot place in context as either background or motivation for a causal inquiry.

BLAU: The descriptive approach is very useful. At its best, it provides an interesting set of facts to be explained, and provides incentives for researchers to generate new ideas, methods, and approaches to explain the facts.

FALK: Good descriptive statistics in the fields you mentioned is a useful first step. They put things into perspective and stimulate new questions and research. When it comes to reporting results from an experiment, descriptive summaries are indispensable. In fact, given the exogeneity of treatments, reporting descriptive statistics by treatment allows already causal inferences. This is of course not true for field

data and here reporting descriptive statistics is much more prone to be misleading, e.g., comparing a particular outcome across countries.

ROBIN: Can one think about social reality without knowing the facts? If a descriptive paper establishes new "significant" facts, to speak like Max Weber, this can be very interesting.

TABER: At some level all empirical work is really descriptive. It is really a question of what kind of filter you use when transposing the data from raw form into something summarized in tables in the paper. I don't think there is any question that descriptive work is extremely useful. For example, even for very structural work the main goal is to try to understand what is happening in the data. The first thing you have to do in such a case is a simple descriptive study to understand the basic data. Ultimately, though, one would hope this is just a first step and further work will try to understand what is driving the basic numbers (either within a given paper or within a literature).

3.2. What do you think of other approaches? For instance, natural experiments versus structural identification is seen as a strong divide by many. Could you give your views on the relative interest for policy of the type of questions posed in treatment effects estimation and in structural estimation? Do not hesitate to be specific and use examples.

ANGRIST: Here is the litmus test in my view: applied structural empirical papers – even the most celebrated – rarely seem to be remembered because of their findings. Structural work seems to be mostly about methods. The big structural hits are often said to be *making progress* or *showing how* to do something, usually something econometrically difficult like estimation of a dynamic multinomial model of something. In Industrial Organization (IO), for example, a hopelessly structural field, some of the big applied papers are about cars or breakfast cereal. Is this work remembered for what the guys who wrote it found or how they did it? Of course, it is easy to beat up on IO. But in this exchange, below, Robin refers to the Keane and Wolpin (1997) paper as being important because it shows "how useful dynamic discrete choice models could be". To be fair, the Keane and Wolpin paper concludes with a brief simulation of the answer to a simple and interesting causal question –the effect of a college tuition subsidy. But that is not what it is usually cited for. It is art for art's sake: The main achievement of this paper, according to most of those referencing it and the authors' own emphasis, is the estimation of a discrete-choice dynamic program-

ming model. As far as substance goes, it does not seem to be meant to be taken seriously. The authors' standard of success is goodness-of-fit. But without some simple alternative benchmark, goodness-of-fit is not worth much –it is just an R-square. And the identifying assumptions used for causal inference are very strong given the authors' interest in allowing for so much endogenous behaviour (e.g., exogenous high school dropout), and jumbled up with all the behavioural assumptions they need for extrapolation. There is no real attempt to assess or justify the causal inferences in this paper.

On the flip side, good natural experiments papers are cited for their findings as well as for the cleverness of the identification strategy. For example, Card's Mariel boatlift paper, my Maimonides Rule paper with Lavy, Alan Krueger's study of class size, and my papers on schooling with Krueger are cited partly for their identification strategies. But because the identification is transparent, the numbers in these papers also became entries in the catalogue of what we know about labor markets and education production. I think it is extremely telling that, in the natural experiments world, when somebody's numbers are called into question in a replication study (as has happened recently to Steve Levitt and Caroline Hoxby), it is big news. I even take some pride in John Bound's critique of the quarter-of-birth estimates in my work with Alan Krueger. One summer at the NBER meetings, we had an intense (for us!) give and take devoted mostly to the substantive question of whether Angrist and Krueger (1991) were really right to conclude that there is not much ability bias in OLS estimates of the returns to schooling. I just do not see how the traditional structural agenda is similarly building up a body of useful empirical findings that are being taken equally seriously.

Another issue is that while structural models are often meant to provide a more general analysis than the causal/reduced-form analyses that I favour, in practice the parameters in structural work are usually highly specific to the model and methods in a particular paper. So it is not clear what good it would do to put the resulting estimates in the empirical catalogue anyway. Going back to the problem of tuition subsidies addressed by Keane and Wolpin, the best evidence –laid out in papers by Sue Dynarski and Tom Kane, among others– comes from direct attacks and careful constructions of the case for causal inference using actual variation in subsidy rates.

BLAU: My inclination is to derive an econometric model from a theoretical model, so that I know how to interpret the parameters I estimate. This does not always mean imposing the structure of the theory on the data, although sometimes that is useful. Rather, the approach can suggest how the interpretation of a parameter estimate depends on what else is controlled in the analysis, and what to control for and not to control for in order to obtain a parameter estimate with a useful interpretation. Often, the natural experimental approach does not provide a clear economic interpretation of the parameter estimate of interest. Calling a parameter estimate a *causal effect* does not seem very enlightening to me. Nevertheless, some natural experiments are quite interesting and good papers have been written with this approach. The best practitioners of this approach take exceptional care to do specification checks to verify the identifying assumptions, which can make their findings quite persuasive. The structural approach has the advantage of readily interpretable parameters, at the cost of much stronger and often unverifiable assumptions. Both approaches are useful, and they should be viewed as complementary, not competing. An illustration of how theory can be helpful in specification and interpretation of a simple econometric model: eligibility for and generosity of child care subsidies vary across states in the US, but it turns out that this variation has little impact on take up and use of subsidies. This is because the child-care subsidy program is severely under-funded, and there are enough funds to serve only about 15% of eligible children. Hence subsidy funds are rationed, and modeling the nature of the rationing process provided useful ideas about how to identify the effects of child care subsidies on employment and related outcomes. I think Josh Angrist's claim that structural work is mostly about methods, while the empirical results that people remember are mostly from simple approaches, has an element of truth but is exaggerated. The Rust-Phelan paper on retirement in *Econometrica* is known for its finding that interactions between employer-provided health insurance and Medicare and Social Security policy can explain variation in retirement timing. The Postel-Vinay-Robin paper on the French labor market in *Econometrica* is known for its finding that search frictions account for the majority of wage variation, i.e. there is substantial within-firm wage variation for workers of similar productivity. And there are well-known examples of papers using simple methods that are remembered mostly for the flaws in the methods, for example the Card-Krueger *AER* paper on the minimum wage.

FALK: Without doubt, all methodological approaches have specific advantages and disadvantages. Risking banality I would therefore say that in general we should view them as complements rather than substitutes. And: the most appropriate approach depends always on the research question at hand.

ROBIN: Science is about understanding facts and mechanisms. Mathematics is not a science because there are no mathematical facts. Economics is about establishing and explaining economic facts (I suppose that some would like to say *all social facts*). Economics is thus not different from physics. Physics has a huge capacity for controlling experiments but not always. Astrophysics, for example, must deduce mechanisms from sometimes very indirect observation. Like astrophysics, economics is a social science with little capacity for controlling experiments.

The more or less recent literature on experimental econometrics is all about the identification of causal effects from quasi-experimental data. The relevant field for *treatment effects* or *natural experiments* is indeed more generally the statistical theory of semi-parametric or non-parametric identification. The literature on *treatment effects* or *natural experiments* is an exemplary case of empirical analyses playing a very important role in fostering theoretical research in statistics or econometrics.

Now, what about structural empirical econometrics? First, one may want to treat economic theories seriously and test them on actual data. My work on equilibrium search model is essentially driven by this motivation. It tries to answer the question: can we understand the determinants of wage distributions? There is a very widely spread representation that any economic fact can potentially be explained by many of alternative theories. This is certainly true but I contest the fact that it would be easy to design coherent economic theories fitting data well. And if two researchers come up with two different theories for the same series of facts, I think this is interesting. By studying how both models differ, we can imagine new surveys or new ways of assembling data to produce falsifying evidence.

I think that our approach of empirical econometrics is too much *instrumental*. If there is no policy analysis in a paper, then it would be worth nothing. I do not think so. Constructing a coherent theoretical description of data can be very useful, if only as a step toward the

construction of a more evolved model allowing for policy analysis. On the other hand, a structural economic model very often offers much more scope for policy analysis than the very limited estimation of one single policy parameter.

Now, this way of opposing structural empirical econometrics and papers interested in producing a consistent estimation of one single policy parameter is both wrong and counterproductive. If you can define a policy parameter, this means that you have built a structural model. At the other end, so called structural models also contain parts which are reduced forms.

This being said, why does such a distinction exist? I think that this is for two reasons. First, there is often the sentiment that the added complications of the structural models are not really useful. Fair enough. Second, so-called structural models are often too complex for a discussion of identification to be completely convincing. This sentiment is reinforced by the fact that, at least in the past, structural empirical papers have made no effort to show that they had the right instruments to control for endogenous selection. As if modelling individual behaviour in a more detailed way rendered the search for instruments unnecessary. I will take only one example. The paper by Keane and Wolpin (1997) is an important paper because it was one of the first papers to show how useful dynamic discrete choice models could be. Now, modelling the dynamics of career choices does not absolve you from instrumenting the fundamentally static initial schooling decision.

I think that people understand all this nowadays and that this opposition between structural empirical econometrics and reduced-form policy analysis will soon disappear.

TABER: I am really bothered by the extent of the natural experiment versus structural identification debate. I agree with David Blau completely when he says that they should be viewed as complements rather than substitutes and I wish everyone viewed it this way. Ultimately to me there is obviously very good structural work and very good *natural experiment* work and there is obviously also poor structural work and poor natural experiment work. For the most part, the problems on which we empirical economists work are very difficult, and to really gain consensus on a problem involves tackling it from a number of different directions. Important contributions have been made using both approaches, and both are needed in the future as well.

That said, in current labor and micro-empirical public economics I think the proportion of structural work versus natural experiments is skewed more towards the natural experiment side than it should be. What bothers me in particular is that I think many people doing natural experiments have only been trained in this approach and do not read and evaluate more complicated structural approaches. I am worried that ultimately this could lead to an even more serious imbalance in the type of work that is done. Generally the nice thing about the natural experiment approach is *internal validity* within the scope of the problem that is being examined. The data experiment is often quite clean (in fact I might even use this as a definition of natural experiment-the gold standard is something that is completely internally valid). However, ultimately as economists we want to address problems for which we cannot find a natural experiment. That is, we need to worry about *external validity*. Taking results from natural experiments and applying them to policy requires making structural assumptions (in fact I might even use this as the definition of a structural assumption-the idea of assuming that a parameter is policy invariant is really a way of saying that it is externally valid). I get particularly bothered by papers that claim not to be structural, but then perform a back of the envelope calculation or even worse make strong policy predictions at the end of the paper. What bothers me about this type of claim is that the authors are trying to have it both ways-claim not to be structural so that they do not have to defend structural assumptions-but then make policy predictions which are only valid under implicit structural assumptions. If the difference between structural and natural experiments is that in structural work you make your assumptions explicit while in natural experiments you leave them implicit-then I am a very strong supporter of structural work. One can even use natural experiment type methods, but still be explicit about the type of assumptions that need to be made to justify the conclusions of the paper. I don't see how one could possibly believe that not writing down a set of assumptions that justify external validity is better than writing them down.

Another aspect of the debate that bothers me is that some natural experiment type researchers will often dismiss structural work as *identification by functional form.* I completely agree that there are examples of poor structural work in which a model is fit without giving proper respect to the data. However, this is not at all a necessary feature of all structural work. Additive separable time and state effects is a very

strong functional form assumption, so *difference in differences* models are an example of something that is identified by functional form. Just because it is the same functional form assumption that a lot of other people are using does not mean that it is not a strong assumption. An advantage of a difference in differences approach is that the map between the data and the numbers in the table is usually more transparent than in most structural approaches. However, this does not mean that the assumptions that justify it are any weaker. Ultimately we need to make strong assumptions in order to identify parameters, but one would hope that the lessons are not sensitive to the functional form assumptions that we make. This is precisely why researchers should be using a variety of tools to attack the same problem rather than arguing about what is wrong and what is right.

3.3. What is the importance of general equilibrium (GE) effects in the evaluation of microeconomic policies?

ANGRIST: For most of the stuff I work on, GE effects are second order. But sometimes I study them. Two examples are my paper on the returns to schooling in the West Bank and Gaza Strip and my paper on human capital externalities with Daron Acemoglu. Both of these papers show that a causal/reduced-form framework can be used to study GE effects. Another example that makes this same point is the Women, War, and Wages paper by my colleagues and former student Acemoglu, Autor, and Lyle, which uses a natural experiment to estimate structural GE parameters. Esther Duflo's thesis on school expansion in Indonesia was also in this vein.

BLAU: The importance of general equilibrium effects is problem-specific. Such effects are very hard to deal with using micro data.

FALK: Experimental approaches are almost exclusively confined to partial equilibrium effects. There are a only a few exceptions where in the lab several interdependent markets are studied. I think it is fair to say that lab experiments are extremely valuable for the understanding of individual choice behavior, strategic interaction, or the study of preferences, motivation and bounded rationality but that they are of limited use for the quantitative evaluation of policies and their general equilibrium effects. An interesting exception is the work by Riedl and van Winden (2003) who studied tax policies in an experimental general equilibrium model. This work was commissioned by the Ministry of Finance.

ROBIN: We do not know yet very well the answer to that question because GE models have not been very widely used in empirical microeconometrics yet. But I think GE models will become very important in the future.

TABER: Microeconometric methods have almost always completely ignored general equilibrium effects. For some policies-maybe even most policies that we focus on, this is probably not a big deal. One would hope that it is not a big deal in any policies, but unfortunately that does not appear to be the case. In my work with Heckman and Lochner (1998b) we examine the effects of a tuition subsidy and find that conventional micro estimates are off by an order of magnitude when one accounts for equilibrium effects. I would not argue that every micro-empirical study should incorporate equilibrium effects, however I think microeconomists should worry about these effects much more than they currently do. There is also a bit of a semantic issue here. I think few would argue that general equilibrium doesn't play an important role in macroeconomic policies. Presumably microeconomists worry about macro policies as well. For example understanding the intertemporal elasticity of labour supply plays a key role in macro models of the business cycle. Labour economists should certainly worry about this.

3.4. More precisely, do you use several approaches?

ANGRIST: I'm not eclectic.

BLAU: I do use several approaches, for the reasons discussed above.

FALK: I use game theoretic models to derive behavioral predictions for my experiments. Another method I often use in combination with experiments is running questionnaires. They help to better interpret and understand the subjects' behaviour in an experiment.

ROBIN: In a given paper, one approach may dominate the others, but I think I made it clear that I was agnostic.

TABER: I hope I made this clear above. I am a strong believer in using multiple approaches on the same problem. I think my *Journal of Political Economy* paper with Cameron (2004) provides a nice example of the way that I think empirical work should be done. We use the same basic idea for identification using several different methods. Some approaches use stronger assumptions than others, but allow answers to broader questions. I think I would characterize my work more as

typically using a lot of econometrics. This is less a belief about the right way to do things, but more about where I think my comparative advantage lies.

4. Best practice examples

4.1. Can you give us your (two) favourite empirical papers (not written by you)? What is special about them?

ANGRIST: Two papers that influenced me greatly are:

Ashenfelter, O. (1983): "Determining participation in income-tested social programs", *Journal of The American Statistical Association* 78, pp. 517-25.

It uses the results of a randomized trial to contrast a mechanical model of eligibility for Negative Income Tax payments with a (slightly) more elaborate behavioural model. The paper is elegant and convincing, with clear findings that provide an important cautionary note for anyone interested in transfer programs.

Lalonde, R. (1986): "Evaluating the econometric evaluations of training program with experimental data", *American Economic Review* 76, pp. 604-20.

It contrasts randomized and observational evaluations of training programs. This was a watershed in social science that helped change the applied micro research agenda and ultimately affected funding priorities.

BLAU: My two favorite empirical papers:

Heckman, J. and G. Sedlacek, "Heterogeneity, aggregation and market wage functions: An empirical model of self-selection in the labor market", *Journal of Political Economy* 93, pp. 1077-1125.

This was one of the first efforts to take the Roy model seriously as a framework for understanding data on labor market earnings and allocation of labor across sectors. The issues examined in the paper are of fundamental importance in economics. The paper has a very solid theoretical foundation, an empirical framework based closely on the theory, thoroughly exploits the available data, and is based on extensive effort to find a specification that fits the data well. It is a very innovative paper that has been highly influential in the 20 years

since it was published, and still provides a starting point for thinking about the wage structure in equilibrium.

Rust, J. and C. Phelan, "How Social Security and Medicare affect retirement behavior in a world of incomplete markets", *Econometrica* 65, pp. 781-832.

I was inspired by this article to work on the substantive issue analyzed in the article and to invest heavily in learning the methods used. Before Rust and Phelan, there was no fully worked-out and estimated dynamic structural retirement model that addressed interesting questions. The question of interest to Rust and Phelan is why do people in the US retire mainly at ages 62 and 65? The obvious answer is Social Security and Medicare incentives: 62 is the first age at which a retirement benefit is available, and 65 is the age at which Medicare is available. However, to demonstrate this empirically is difficult because these features of the Social Security system have been unchanged for many decades. That is why a structural analysis is especially useful in this case. The article combines a clear statement of the model together with very extensive work on the data, and a convincing set of findings.

FALK: This is a tough question! Before I answer it, let me first reduce the possible set to papers that report experiments. Two papers that I really like are:

Fehr, E. and S. Gächter (2000): "Cooperation and punishment in public goods experiments", *American Economic Review* 90, pp. 980-994.

This paper studies the power of social preferences for the enforcement of cooperation. The experiment is a simple two-stage game. On the first stage subjects can contribute to a linear public good. While contributing is efficient it is a payoff dominant strategy not to contribute at all. On the second stage subjects are informed about the cooperation of the other group members and are given the chance to punish others at a cost. In contrast to the standard model, assuming material self-interest, subjects punish defectors, even though it is costly. The reason is that they are reciprocally motivated, i.e., they reward kind and punish unkind behavior. As a consequence of the reciprocal punishments relatively high cooperation levels can be sustained. The experiment has been replicated in many versions and has helped me to better understand the role of social preferences for solving free-rider problems. It is a path breaking paper.

Gneezy, U. and A. Rustichini (2000), "A fine is a price", *Journal of Legal Studies* 29, pp. 1-17.

This paper is a field experiment and shows that material incentives can backfire. Uri and Aldo study this question in daycare centers in Israel. The problem in many daycare centers is that parents come late to pick up their child. A natural economic solution to this problem is implementing a fine for late coming parents. But would that work? To answer this question, the authors study a control and a treatment group both consisting of several daycare centers. In the control group there are no fines. In the treatment group there are no fines in the first phase; fines are introduced in the second phase and removed in the final third phase. The standard economic model would predict that late comings are not increasing in the second or third phase of the treatment group, but they do! In the treatment group late-comings increased after the introduction of fines and settled at an almost twice as high level as the initial one. Removing the fine did not affect the number of late-comings. The study neatly shows that the psychology of incentives is much more complex than economists often think and that policy advice built on our simplistic models may be severely counterproductive. This influential paper stimulated a lot of debate and inspired many new studies.

ROBIN: No. I learnt from so many papers... But I have the greatest respect for the whole work of Jim Heckman and Dale Mortensen (who is not an econometrician but whose economic theory is so useful to empirical economists –I think that this is because Dale cares about writing economic models which explain real data).

TABER: I am not quite sure how to answer this. To some extent my favorite empirical paper would be the one which changed my view of the state of the world in a positive way (for example a paper showing the decline in AIDS in the US) However, this is more about the data than the methodology. Let me then change the question and instead name the two papers that had the largest influence on me in how I approach empirical work. Put that way I would include econometrics papers along with empirical papers. I think there were really a series of papers by Heckman and a number of other people on the late 1970s and 1980s on heterogeneity, self selection, and identification that had a huge influence on me and my approach. If I were forced to pick two papers, they would be Willis and Rosen (1978) and Heckman and Singer (1984). Much work that I have done has been on schooling

choices with self-selection and returns to schooling. Willis and Rosen (1978) has always been my starting point as an econometric model for thinking about the process. While I have done very little work on duration models per se, the Heckman and Singer (1984), the style of modelling heterogeneity and trying to formally show identification of the structural model is an approach that I try to emulate.

4.2. What is your own best empirical article? Why?

ANGRIST: A paper I am particularly proud of is my study of voluntary military service, published in *Econometrica*, 1998. This paper does not have super-clean identification. But I was deeply committed to it and worked on it longer and harder than any of my other projects, before or since.

BLAU: My choice is "The supply of quality in child care centers", co-authored with Naci Mocan, *Review of Economics and Statistics* (2002, pp. 483-496). I like it because a) doing the research pushed me in new directions, b) the data are unusual and very rich, and as a result I invested a lot of effort in preliminary descriptive and exploratory work, which makes me feel that I understand the data well, c) there is a good blend of theory and econometrics, with the empirical specifications based directly on the theory and therefore easily interpretable, and the estimation methods simple but appropriate, and d) the topic is new, and there has been little previous work on it.

FALK: My best article is not written yet... But up to date I think the papers "Contractual incompleteness and the nature of market interactions" (with Martin Brown and Ernst Fehr) and "The hidden cost of control" (with Michael Kosfeld) excite me the most. In the first paper we provide evidence that long-term relationships between trading parties emerge endogenously in the absence of third party enforcement of contracts and are associated with a fundamental change in the nature of market interactions. We show that without third party enforcement, the vast majority of trades are initiated with private offers and the parties share the gains from trade equally. Low effort or bad quality is penalized by the termination of the relationship, wielding a powerful effect on contract enforcement. Successful long-term relations exhibit generous rent sharing and high effort (quality) from the very beginning of the relationship. In the absence of third-party enforcement, markets resemble a collection of bilateral trading islands rather than a competitive market. If contracts are third party enforceable, rent sharing

and long-term relations are absent and the vast majority of trades are initiated with public offers. Most trades take place in one-shot transactions and the contracting parties are indifferent with regard to the identity of their trading partner.

In my paper with Michael on trust and control, we show that controlling signals distrust and undermines motivation. We study an extremely simple experimental principal-agent game, where the principal decides whether he controls the agent by implementing a minimum performance requirement before the agent chooses a productive activity. Our main finding is that a principal's decision to control has a negative impact on the agent's motivation. While there is substantial individual heterogeneity among agents, most agents reduce their performance as a response to the principals' controlling decision. The majority of the principals seem to anticipate the hidden costs of control and decide not to control. In several treatments we vary the enforceable level of control and show that control has a non-monotonic effect on the principal's payoff. In a variant of our main treatment principals can also set wages. In this gift-exchange game control partly crowds out agents' reciprocity. The economic importance and possible applications of our experimental results are further illustrated by a questionnaire study, which reveals hidden costs of control in various real-life labor scenarios. We also explore possible reasons for the existence of hidden costs of control. Agents correctly believe that principals who control expect to get less than those who don't. When asked for their emotional perception of control, most agents who react negatively say that they perceive the controlling decision as a signal of distrust and a limitation of their choice autonomy.

ROBIN: This is always the last one. I believe in human capital accumulation.

TABER: I think my project with Heckman and Lochner on estimating and simulating general equilibrium effects is my best work. I would really point to two closely related papers (Heckman, Lochner, and Taber 1998a) which estimates the model and shows the importance of accounting for human capital accumulation in wage growth and (Heckman, Lochner, and Taber 1998b) which shows the importance of general equilibrium effects in micro problems. I think when combined, the papers show a) that accounting for general equilibrium effects in

micro empirical work can be very important, and b) that while difficult it is feasible.

Complementary References

Angrist, J. and A. Krueger (1991): "Does compulsory school attendance affect schooling and earnings?", *Quarterly Journal of Economics* 106, pp. 979-1014.

Brown, M., A. Falk and E. Fehr (2004): "Relational contracts and the nature of market interactions", *Econometrica* 72, pp. 747-780.

Cameron, S. and C. Taber (2004): "Borrowing constraints and the returns to schooling", *Journal of Political Economy* 112, pp. 132-182.

Dohmen, T., A. Falk, D. Huffman, U. Sunde, J. Schupp and G.G. Wagner (2005): "Individual risk attitudes: New evidence from a large, representative, experimentally-validated survey", IZA Discussion Paper 1730.

Falk, A. (2004): "Charitable giving as a gift exchange: Evidence from a field experiment", IZA Discussion Paper 1148.

Falk, A. and M. Kosfeld (2004): "Distrust - the hidden cost of control", IZA Discussion Paper 1203.

Heckman, J., L. Lochner, and C. Taber (1998a): "Explaining rising wage inequality: Explorations with a dynamic general equilibrium model of labor earnings with heterogeneous agents", *Review of Economic Dynamics* 1, pp. 1-58.

Heckman, J., L. Lochner, and C. Taber (1998b): "General equilibrium treatment effects: A study of tuition policy", *American Economic Review* 88, pp. 381-386.

Heckman, J., and B. Singer (1984): "A method for minimizing the impact of distributional assumptions in economic models for duration data", *Econometrica* 52, pp. 271-320.

Keane, M. and K. Wolpin: "The career decisions of young men", *Journal of Political Economy* 105, pp. 473-522.

List, J.A. and S.D. Levitt: "What do laboratory experiments tell us about the real world?", University of Chicago Working Paper.

Riedl, A. and F. van Winden (2003): "Input versus output taxation in an experimental international economy", CREED Discussion Paper.

Willis, R. and S. Rosen (1979): "Education and self-selection", *Journal of Political Economy* 87, S7-S36.

Resumen

Este artículo presenta una discusión entre economistas de primera línea acerca de cómo realizar investigación empírica en economía. Los participantes presentan los motivos que les llevan a elegir un proyecto, la construcción de bases de datos, los métodos que emplean, el papel de la teoría y su visión sobre los principales enfoques empíricos alternativos. El artículo finaliza con una discusión sobre un conjunto de artículos que representan modelos a seguir en la investigación aplicada.

Palabras clave: Investigación empírica, métodos econométricos.

Recepción del original, diciembre de 2005
Versión final, febrero de 2006