DISCUSSION PAPER SERIES

IZA DP No. 2700

Policy Evaluation and Economic Policy Advice

Christoph M. Schmidt

March 2007

Forschungsinstitut
zur Zukunft der Arbeit
Institute for the Study
of Labor

Policy Evaluation and Economic Policy Advice

Christoph M. Schmidt
RWI Essen, Ruhr University Bochum and IZA

Discussion Paper No. 2700
March 2007

IZA

P.O. Box 7240
53072 Bonn
Germany

Phone: +49-228-3894-0
Fax: +49-228-3894-180
E-mail: iza@iza.org

IZA Discussion Paper No. 2700
March 2007

ABSTRACT

Policy Evaluation and Economic Policy Advice*

Arguably, one of the most important developments in the field of applied economics during the last decades has been the emergence of systematic policy evaluation, with its distinct focus on the establishment of causality. By contrast to the natural sciences, the objects of our scientific interest typically exert some influence on their treatment status under the policy to be evaluated and on their economic outcomes. Thus, economic policy advice can only be successful, if it is based on an appropriate study design, experimental or observational. It will strive in societies that provide liberal access to data, accept the merits of randomized assignment and guard the independence of research institutions.

JEL Classification: A11, C01, H50

Keywords: policy evaluation, applied economics, causality, policy advice

Corresponding author:

Christoph M. Schmidt
RWI Essen
Hohenzollernstraße 1-3
45128 Essen
Germany
E-mail: schmidt@rwi-essen.de

* I am grateful to Jochen Kluve, Joachim Schmidt and Marcus Tamm for their comments and to Claudia Lohkamp for her support in preparing the manuscript.

1. The (German) Market for Economic Policy Advice

When speculating about the way in which economic policy advice will be conducted in the future, and about the methodological basis on which it will rest, one has to realize first that very little is known about the actual extent to which applied economic research affects public policy. This is true both within and across countries: while it seems that different countries have established their own individual modes of conducting research and providing policy advice, the experiences made with these approaches are usually not compared systematically. This is all the more remarkable, as academic research in economics is conducted in an integrated world market. Nonetheless, in some countries (e.g. the United States) policy makers and administrators tend to demand sophisticated, methodologically rigorous economic research as a basis for policy advice, while in other countries (e.g. Germany) they even struggle with researchers over side issues such as data confidentiality and the ethical basis of social experimentation.

To leave large parts of the potential for economic policy research untapped most likely implies substantial costs, which ultimately arise in the form of implementing and retaining ineffective and wasteful policy measures. This research potential, moreover, has grown tremendously, as perhaps the most important developments in the field of applied economics during the last decades have been the emergence of systematic policy evaluation, and an increasing awareness of the pitfalls associated with inadequate empirical research strategies. These developments bring into focus the assessment of causal effects: Whereas it is often the task of economists to characterize certain economic phenomena in a thorough descriptive analysis, or to provide a reliable forecast of future developments, the key aspect of policy evaluation is the establishment of causality. This is far from being a trivial task, due to a fundamental requirement faced by most applied economic research, i.e. to answer research questions outside of a comfortable and perfectly controlled laboratory environment.

In Germany, the emergence of systematic policy evaluation constitutes only part of a whole range of recent changes to the market for economic policy advice[2]. In fact, for a number of reasons, both the demand side and the supply side of applied econometric research and consulting have changed tremendously during the last decade. Certainly, the economic problems to be addressed have become more complex with increasingly open and

[2] The role of the “large” German economic research institutes gathered in the Leibniz Association (WGL) is discussed in Schmidt (2006).

interdependent markets, with faster production cycles and more immediate diffusion of knowledge, and with encompassing demographic and societal change,. Most importantly, though, the increasing awareness of the limitations of government interventions along with tighter public budgets have generated a growing pressure to thoroughly evaluate the effectiveness of policy measures. Being held more accountable than in the past, politicians and administrators often request quantitative and reliable assessments directly after or even before the implementation of policy measures, rather than abstract long-run advice.

Parallel to these changes on the demand side of the market, the supply side has experienced dramatic alterations as well. While in the past the division of labor between statistical offices (data collection and provision), research institutes (descriptive analyses and extrapolations based on the "current fringe" of data), and the universities' economics faculties (economic research) seemed to be sensible, this organizational arrangement has been fundamentally challenged. First, this strict division of labor is quite counterproductive, since good policy advice requires a study design which is deeply rooted in economic and econometric research. Second, as evidenced in the past by the often moderate academic quality of the work of economic research institutes and by the limited relevance of much of the university research, what seemed at the time as a sensible division of labor has led to sever the link between research and policy advice.

Third, the almost complete overhaul of organizational structures, methodological approaches, and academic aspirations experienced by Germany's economic research institutes has pushed them to the forefront of applied economic research in the country. It is certainly not a mere coincidence that researchers from these institutes have accounted for a good chunk of those publications in the top journals in economics which have been written by German scholars in recent years. At the same time, fourth, other players have entered the market, generating a variegated set of suppliers of economic policy advice. These new competitors are mainly university-based researchers, incentivized to address applied "real-world" problems by shrinking financial resources, and private consulting companies whose intellectual base has steadily improved, augmenting their hands-on approach to finding solutions to problems of organization and implementation.

Thus, overall, it is the increase in competition on the supply side that has prevented individual suppliers to lean back and simply enjoy the increasing demand for economic policy advice, and that has put weak or intellectually stagnant players under tremendous pressure, hence improving the overall quality of work. These developments were supported by advances in

technology and recent progress in empirical methodology. On the technical side, advances comprise the remarkable increase in the access to comprehensive micro-level data, even in continental Europe, and the surge in computing power available for handling large and complex data bases. More importantly, though, in recent years applied research in econometrics and statistics has emphasized – under the heading "identification" – the quest for both the potential and the limitations of (non-experimental) empirical research, rather than an ever increasing level of technical sophistication[3]. This has pushed economics forward as an empirical science.

Against this background, the issue of identification is also a key part of this paper, as it focuses on the evaluation of economic policy measures. The analysis of causality, in particular the assessment of policy interventions with respect to their effects, is one of three fundamental tasks of empirical research in economics, and perhaps its hardest intellectual challenge. The two other tasks are descriptive analysis and forecasting. The specific objective of evaluation studies in economics is the isolation of the effects of the policy intervention under study to the best extent possible from the impact of all other aspects of the economic environment. Applied research in economics has made tremendous progress in reaching this objective and evaluations of labor market interventions have played a particularly important role in carrying this progress forward. It is now up to the demand side to make full use of this potential for the design of better economic policy.

In the next section, I will discuss in intuitive terms the aspects that make identification problems such a tremendous intellectual challenge, characterizing them as problems of limited observability (i.e. data availability). The third sectiion discusses suggestions how the current state of knowledge can be utilized to offer convincing solutions to this problem

2. Identification Problems in Economic Analysis

The key to good empirical research in economics is finding the right balance between quantitative skill and economic expertise. Quite fundamentally, proficiency in mathematical statistics is insufficient if it is not accompanied by a good dose of economic competence, such as identifying "what is the object of interest?". For instance, a particular conditional expectation (the mean of Y given X) might be a prime candidate for empirical economic analysis, while another (the mean of X given Y) is without any economic content. Yet, given the data the mathematical statistician will be able to extract an estimate of both entities.

[3] A seminal contribution to the issue of identification is Manski (1995).

Similarly, and of equally fundamental importance, before starting the analysis the researcher needs to assemble everything she firmly believes to hold true, so firmly that it will not have to be questioned throughout the analysis ("what do I know already?").

If this *a priori* knowledge were to entail every aspect of the phenomenon (in the terminology of mathematical statistics: the complete distribution), no empirical study would be necessary. At the other end of the spectrum, if no baseline knowledge existed to rest the analysis on, an empirical study would be impossible. Since empirical research is the attempt to learn something about the properties of a probability distribution (or "population") from the repeated observation of realizations drawn from the same underlying distribution (or "sample"), a minimum requirement is the assertion that indeed (parts of) the data represent draws from the same population. Since the ultimate truth about such aspects can never revealed with certainty, from the perspective of the researcher this assertion is in essence an assumption that will not be questioned any further, a so-called *identification assumption*.

Typically, maintained assumptions of this sort need to go beyond this minimum requirement, dictated by the wealth of data. It would make little sense, for instance, to attempt the estimation of a highly non-linear relation, if only a few data points in the sample allowed the researcher to trace it across the relevant range of potential observations. In such cases, resorting to the analysis of a restrictive model (i.e. a model whose properties are expressed in a limited number of parameters), for instance a linear regression model, is the only choice. Maintaining the identifying assumption that the sample represents a population for which this restrictive model unquestionably holds, reduces the quest for further knowledge to the estimation of the model parameters.

Empirical research cannot extract more pieces of information from the data than the data provide – it needs so-called degrees of freedom: if several data points "voluntarily" demonstrate a similar realization, although they are free to differ widely, then this can be taken as a firm indication for the relevance of this similarity. These matters of statistical inference are well understood. In particular, we know that the confidence about the validity of approximations to population properties rises with sample size – holding the identification assumptions fixed. They are also the basis for systematic hypothesis testing – again, maintaining the identification assumptions. Thus, at the beginning of any empirical study one must ask what sort of questions the available sample does allow asking ("what can I learn with some confidence?"). In theory, sample sizes are irrelevant for the formulation of identification assumptions, in practical applications they are not.

In addition to this trade-off between identification assumptions and questions left open to statistical inference, many phenomena do not become observable more precisely as the sample size increases. For instance, the conditional expectation of a random variable Y (the "outcome") given a random variable *X* (the "policy") might not depend at all causally on the specific value of *X*, but the realizations of both *Y* and *X* might reflect the specific values of yet another random variable *Z*. Any analysis ignoring *Z* would then lead to erroneous conclusions about the causality from *X* to *Y*. If the sample does not contain information on *Z*, then collecting more and more realizations of *Y* and *X* will not help. Thus, only an identification assumption ruling out the relevance of any other conditioning factor *Z* allows for the causal interpretation the researcher is aiming at. Again, there is a trade-off involved: Imposing a correct identification assumption allows the researcher to extract meaningful insights from the data, but the more stringent these restrictions have to be in order to proceed, the higher is the risk to derive completely erroneous conclusions.

Since they are the prerequisite for causal inference, identification assumptions, such as ruling out any contamination by a third factor *Z*, cannot be questioned any further throughout the analysis. Consequently, they can never be subjected to any statistical test either. Instead, the researcher has to ask herself at the outset whether she can convincingly construct an answer to the underlying counterfactual question: "What would have happened to the outcome of interest *Y*, if – for the same observation unit – the realization of the policy variable *X* had been different from what actually happened?" Since it is impossible to observe different scenarios for the same observation unit in the data, it is clear that somehow the answer to this question must be derived from different experiences of different observation units.

This might or might not be possible, however, depending on the specific situation at hand. In principle, any economic researcher would cherish a situation in which the policies under study were applied exogenously to the individuals (households, enterprises, regions, etc.) whose outcomes (employment, profits, unemployment rates, etc.) those programs intend to alter. That is to say that neither the units under treatment nor the program administrators implementing the policy were to base policy participation decisions on characteristics of candidate participants. Then the difference between the typical outcome for units participating in the program and units who do not participate will provide a convincing estimate of the program effect – one would truly compare the comparable. In practice, one way to meet this requirement is the randomized assignment of units to treatment and control groups, respectively. Many applied studies need to proceed outside of a controlled experiment,

though, confronting so-called *observational* data. In fact, in certain cases an intelligent study design can in principle produce a non-experimental estimate of the program effect that is similarly convincing as the experimental ideal.

One obvious way to generate empirical evidence on program effectiveness is to use all observable characteristics of the individual program participants and non-participants to stratify the sample, and then to conduct the analysis within homogeneous population strata. Yet, in practical applications, this approach is not necessarily successful: The most important principle in the evaluation of social policies is that the objects of our scientific interest, namely the individuals addressed by the policy, have their own minds and consequently make their own decisions. These decisions necessarily influence the state of (material) welfare that people experience without the policy, as well as the effect that it exerts in case of its implementation. Thus, while in their specific lines of research bio-engineers, pharmacologists or physicists might have to analyze very complex mechanisms, their objects of interest, say plant or animal cells, human organs, or photovoltaic sensors, are usually not responsible for their own destiny. By contrast, when assessing if and how a social policy works for the individuals it is targeted at, the social scientist usually needs to understand first why individuals arrived at the state which the policy has been designed to alter. The policy can only be effective, if it manages to change the behavior of its clients in the desired direction.

If people were indeed behaving like mechanical devices, the impulses created by a social policy could be assessed easily. One would simply administer the treatment implemented by the program to some target individuals, while withholding it from others which are used as comparison observations. The treatment would be exogenous in the sense discussed above. In this hypothetical world, individuals would display perfect compliance with their assignment, i.e. target individuals would follow the prescribed treatment and comparisons would stay out of the program or any substitute for it. Given that both groups of people are observed in the same environment, the estimated effect of the social policy could then be read off directly from the difference between the outcomes of the target individuals and their comparisons. This procedure would mimic the research strategy that natural scientists pursue in the laboratory. In the reality addressed by social policies, though, people very well have a say in their own actions.

Specifically, in the social sciences, even people who appear observationally equivalent might be quite different in aspects that directly pertain to the outcome under study, but that remain hidden to the researcher (e.g. motivation, work ethic, etc.). These characteristics might also

influence the compliance with the original assignment into target and comparison groups. In practice, treatment under a social program is an offer that has to be taken up by targeted individuals, and in the absence of access to the social policy comparison individuals might seek out alternative ways to improve their situation. Thus, the individuals choosing to participate in the program might be quite different – as regards characteristics not observed by the researcher – from those who do not make this decision. Taking the latter as a comparison group for the former might therefore be a problematic strategy for estimating the effect of the policy. Basically all recent advances in the econometric literature on program evaluation have aimed at developing methods to properly account for this unobserved heterogeneity, i.e. coming close to estimating the "true" counterfactual outcome.

3. Elements of Rigorous Program Evaluation

Rigorous program evaluation emerged as a major innovation in the implementation of economic and social policy during the second half of the last century[4]. It is now widely recognized that advancing the state of knowledge on which programs have worked in the past, and which have not, enables policy makers and administrators to make informed predictions about outcomes of future interventions, and to design their policies accordingly. Thus, by pointing out the potential and the limitations of systematic evaluation of past programs, economic research has been most valuable to policy. This progress has emerged along three routes, (i) formulating the right questions, (ii) getting access to highly informative data, and (iii) establishing causality.

First, before developing a specific research design, economists tend to ask a series of preliminary questions. One aspect that they emphasize, next to the definition of appropriate units of observation and the choice of relevant outcome measures, is whether the program under scrutiny indeed had any effect on the environment in which these decision units operate. Mandating a change or disbursing money is not the same as actually providing substantive treatment. Only an intervention exerting true change – to individuals' human capital, to firms' labor cost, to regions' infrastructure etc. – can be successful in altering the relevant outcomes. Most importantly, economic policy evaluation concentrates on posing the appropriate counterfactual question, and on deriving a convincing answer from the available data.

[4] A comprehensive overview is given by Heckman et al. (1999); see also, among many others, Schmidt (1999) or Blundell and Costas Dias (2000).

Hence, second, access to informative data is a crucial condition for solid economic policy advice. Most applied problems cannot be discussed exclusively in the form of theoretical models, since their conclusions typically rest on the signs or even precise magnitudes of key parameters that can only be gathered from empirical data. Also other actors – such as the administrators implementing the program, or the statistical offices – cannot generally be relied upon to extract this information for the researchers, since complex empirical problems typically require approaches that are tailor-made to the evaluation study, rather than one-size-fits-all. This fundamental requirement for data to be useable for economic policy advice has led the most successful research institutes today to engage in collecting and linking data, and even providing data to other researchers. Unfortunately, quite different from the US or Scandinavia, it is still comparatively difficult to get access to individual-level data in Germany[5].

Third, and perhaps most importantly, good economic policy advice needs to be based on a solid assessment of causes and effects. To establish causality empirically, the "gold standard" is arguably a randomized controlled trial. Yet, at least in continental Europe it is all but impossible to realize random assignment of treatments[6], as typically the policy side, politicians and administrators alike, are decidedly reluctant to do so. The argument raised most often regards the ethical caveat, i.e. no person should be deliberately withheld from a potentially beneficial treatment. Clearly, this point has more force in the medical context, if e.g. the treatment is a drug against HIV/AIDS. Regarding labor market programs, however, it is unlikely that the outcome without the treatment is detrimental, since in continental Europe there is a comfortable social safety net and negligible levels of absolute poverty[7].

Another potential explanation for the political reluctance towards randomized social experiments is that conducting experiments simply limits the exertion of political influence. It is much more difficult to re-interpret results from social experiments than the results of non-experimental studies. Whereas the latter ones tend to be sensitive with regard to the concrete study design chosen and are therefore always subject to qualifications, results from a social experiment are easy to understand and easy to communicate. Consequently, negative results from a non-experimental study may not prevent the continuation of a program, but if experimental results spell out "failure", this will tend to kill the funding for it.

[5] There has been some progress, though, arising from the introduction of so-called *Forschungsdatenzentren* that have the explicit task of providing researchers with access to selected individual-level data sets.
[6] Comprehensive surveys of the existing literature on the evaluation of labor market interventions in Europe are Kluve and Schmidt (2002) and Kluve et al. (2007).
[7] This might be different for the issue of relative poverty or, rather, income inequality.

Traditionally, in the European Union policy makers displayed a relatively low interest in having their interventions evaluated. This lacking "evaluation culture" is changing gradually, though, and more and more countries have some of their policies evaluated, a process frequently triggered by the European Commission as a requirement on EU-co-financed programs (Kluve et al. 2007). In Germany, the recent reforms of the labor market ("Hartz reforms") were the first policy measures for which the parliament explicitly requested a thorough scientific evaluation of the effects of the reforms. Increasing public pressure has certainly helped in getting this change started, as has the growing understanding of the limitations of traditional research strategies among administrators.

If random assignment is not possible, researchers might find clever ways of emulating a randomized controlled trial *ex post*. This approach exploits the "natural experiment" introduced by events that are arguably exogenous to treatment choice. There has been some inflationary use of this term in the literature, but in a sense, all observational studies need to ascertain that, given that all the observable confounding factors are controlled for in the analysis, assignment to treatment has been random, as it would have been in an experimental study. If that is not the case, all that can be established with certainty in an observational study are correlations. It is then up to the researcher to make the case for identification assumptions that support a causal interpretation of these correlations. Fortunately, we know much better today than in the past how to construct a counterfactual convincingly from the data.

Given these methodological advances, how should economic policy advice be organized?. First of all, to fulfill its potential, economic research needs to be independent of vested interests. In Germany, this is ascertained for many institutions offering economic policy advice, most prominently for the universities and the "large" research institutes gathered in the Leibniz Association (WGL). But independence alone is not enough. Rather, if institutions want to ensure a steady flow of academic talent to applied economic research, both the profession itself and the policy side need to treat publications in refereed journals and good policy advice with equal respect. This is challenging, since the quality of policy advice is often difficult to measure. But only then will they be able to find the right balance between academic research and policy advice. To be successful in the long run, German universities and research institutes need to educate new generations of government administrators. This will improve the communication and the understanding between the policy side and the

research side, and most likely pave the way to better policy advice and improved economic policy.

Often markets change, because new players enter, doing the same things as the incumbents. Yet, they might also change, because advances in technology alter the actions of all players in the market. This is what happened to the market for economic policy advice. Most importantly, increasing awareness of the limitations of empirical research has pushed the emphasis away from complex mathematics towards a careful scrutiny of the underlying study design, that is, towards the formulation of counterfactual questions and the issue of identification. Although the modern methods of economic policy evaluation originate mainly in labor economics, they are applicable more broadly. Recognizing the potential of this different emphasis for improving the evaluation of policy interventions not only in the labor market, but also in health, education, the environment, and basically all other areas, in which governments intervene into the functioning of the market, promises to lead to the design of better policy[8].

[8] Recent discussions of these ideas address, for instance, poverty alleviation in the developing world (Ravallion 2007) or energy policy (Frondel and Schmidt 2006).

References

Blundell, R. and M. Costas Dias (2000), Evaluation Methods for Non-experimental Data, *Fiscal Studies* **21**: 427-468.

Frondel, M. and C. M. Schmidt (2006), The Empirical Assessment of Technology Differences: Comparing the Comparable, *Review of Economics and Statistics* **88**: 186–192.

Heckman, J. J., R. J. LaLonde and J. A. Smith (1999), The Economics and Econometrics of Active Labor Market Programs, in: Ashenfelter, O. and D. Card (eds.), *Handbook of Labor Economics vol. 3*, Amsterdam: North-Holland.

Kluve, J., D. Card, M. Fertig, M. Gora, L. Jacobi, P. Jensen, R. Leetmaa, L. Nima, E. Patacchini, S. Schaffner, C.M. Schmidt, B. van der Klaauw , A. Weber (2007), *Active Labor Market Policy in Europe: Performance and Perspectives*, Springer: Berlin et al.

Kluve, J. and C. M. Schmidt (2002), Can Training and Employment Subsidies Combat European Unemployment?, *Economic Policy* **35**: 411-448.

Manski, C. F. (1995), *Identification Problems in the Social Sciences*. Cambridge, Mass. et al.: Harvard University Press.

Ravallion, M. (2007), Evaluating Anti-Poverty Programs, forthcoming in: Evenson, R. E. and T. P. Schultz (eds.), *Handbook of Development Economics vol. 4*, Amsterdam: North-Holland.

Schmidt, C. M. (1999), Knowing What Works, *IZA Discussion Paper* No. 77.

Schmidt, C. M. (2006), Fokus, Fokus, Fokus? Zur Rolle der außeruniversitären Wirtschaftsforschungsinstitute, *Allgemeines Statistisches Archiv* **90**: 617-622.