Comment on John P. Martin and David Grubb: What works and for whom: A review of OECD countries' experiences with active labour market policies

Anders Björklund[*]

It was a pleasure to read this paper; it provides an informative review of a huge literature, and a careful interpretation of the results. I found it particularly interesting to read the discussion about what programs seem to work and under what circumstances. Furthermore, I also think that the section on "interventions in the unemployment spell" raises questions in the Swedish public policy discussion, that deserve more attention.

My main concern is that the authors do not explain to readers who are skeptical to evaluation research, why they should believe in all reported results. I know that there are many skeptics to evaluation studies among policy makers and officers at important public authorities like the Swedish National Labor Market Board. Despite impressive methodological insights, I, myself, have also become less and less convinced by the results in typical studies of Swedish and European labor market polices.

The main motivation for my skepticism to typical program evaluation studies is that a convincing study must solve the well-known selection problem. The most likely participant in a Swedish program is an unemployed person who considers himself, and/or is considered by program administrators to benefit particularly from participation. Suppose that we attempt to estimate the program impact on such people's future earnings. To do this, we must infer what earnings a person would have had if he had not participated in the program. In a typical study, such inference must be made from earnings data on those unemployed who have not participated. If these non-participants (as well as the program participants) were randomly selected from a pool of applicants to the program, I would find the results quite convincing. In particular, this would be the case if the randomization itself had not affected the way the program is run. But, in

[*] *Anders Björklund is professor of economics at the Swedish Institute of Social Research at the Stockholm University.*

practice, the inference must be made from non-participants who have been selected by the system as it works in practice, i.e. by the complicated interplay between the unemployed person and the program administrators. Thus, the selection problem is a tough one to solve. Furthermore, I think that the burden of the proof falls on the researcher, who should be able to explain to a non-technical layman why the problem has been solved and why the results are compelling.

Those of us who have followed evaluation research for a while can witness that researchers have generally been quite excited by the most recent estimation techniques for solving the selection problem. Slightly over 20 years ago, when I entered this field of research, panel data had become available. Then, we were quite optimistic about the prospects of solving the selection problem by exploiting the before-after dimension of panel data on participants and non-participants. A few years later, so-called sample selection techniques became fashionable and many of us thought they were a panacea for evaluation research. Nowadays, various matching techniques seem to be the most popular ones; "non-parametric propensity score matching" seems to be among the most fancy techniques at present.

So why should a person who has followed this development over a couple of decades suddenly believe in results such as those summarized by Martin and Grubb? A few results, mainly US ones, stem from successfully implemented randomized experiments so I have no problem with these. But what about the other studies, and almost all European ones?

One argument could be that nowadays, researchers have several techniques for non-experimental studies at their disposal, and they can determine which one best fits a specific evaluation problem. This argument was stressed by Heckman and Hotz (1989), who argued that the annoying variation in estimated program effects using alternative non-experimental estimators could be reduced by employing their proposed specification tests. The enthusiasm over these tests disappeared quite quickly, however. It is somewhat comforting, though, that Smith (2002) very recently, and drawing on one Swedish study (Regnér, 2002) and one Norwegian study (Raaum and Torp, 2002), concludes that these tests have some value in reducing the confusion arising when alternative seemingly reasonable estimators yield markedly different results. Thus, Smith concludes that these specification tests deserve more attention that they have received in the literature.

My impression, however, is that a great deal of uncertainty remains in these studies, even after applying the specification tests.

There is one approach in the evaluation literature that really makes an impression on me. What I have in mind are studies like Heckman et al. (1998), which investigate the performance of non-experimental estimators by comparing them with results from randomized experiments. More specifically, they use a comparison group generated from the selection that takes place in the real world and investigate if it is possible to replicate the estimated effects generated by a randomly selected control group. With this research approach, they could document the importance of data quality in general, as well as the importance of information on local labor market conditions to achieve results close to those from the experiment. Further, they found that conventional versions of matching, sample selection and panel-data techniques estimate substantial biases, whereas "non-parametric" versions perform much better. Unfortunately, most previous studies used the "parametric" versions.

Although the distinction between non-experimental estimators that perform well and those that do not is quite subtle, it lends some credibility to US evaluation studies that use methods having survived such an examination. But, in my view, the problem is that there are no strong reasons to believe that estimators that seem to have done a good job in the US should do the same in Europe. The labor markets, as well as the selection processes into labor market programs, are different.

The experience from studies like Heckman et al. (1998) demonstrates that randomized experiments are useful in two ways. First, they provide compelling results per se. Second, they can be used to evaluate the performance of evaluation techniques that must be used when randomization is not feasible. Unfortunately, the US tradition to evaluate new programs by randomized experiments is very seldom followed in Europe. The only Swedish experimental study was done in 1974, i.e. 28 years ago.[1] It is also unfortunate that it has not been possible to use the data from that study to evaluate the performance of non-experimental techniques.

Since most European and all Swedish studies have used non-experimental techniques that have not been evaluated with the same

[1] See Björklund and Regnér (1996) for an examination of US and European randomized experiments on labor market policy through the mid-1990s.

scrutiny as the US ones, I am much less convinced by the findings from this side of the Atlantic. I hope that Martin and Grubb will use their powerful platform at the OECD to stress the importance of randomized experiments.

Another kind of evaluational study that could convince a skeptic like me—but maybe not the skeptics among policy makers—is the one using the instrumental variable technique. This means that only the variation in program participation generated by a known and arguably exogenous source—the instrument—is used to estimate the program effect. In recent years, this technique has become quite popular for estimating the return to schooling. Krueger and Lindahl (1999) offer a long list of such studies. The instruments generating the useful variation in choice of schooling are often characterized as natural experiments; quarter of birth, distance to college, and policy interventions that only affect some individuals are typical examples. To me, it is a mystery that so many convincing instruments have been found in the study of the return to general schooling, but none in the study of labor market programs. Hopefully, there are a number of useful instruments—or natural experiments—waiting to be detected in the study of European labor market programs.

References

Björklund, A. and Regnér, H. (1996), Experimental evaluations of European labour market policy, in G. Schmid et al. (eds), International Handbook of Labour Market Policy and Labour Market Evaluation, Edward Elgar, Cheltenham.

Heckman, J. and Hotz. J. (1989), Choosing between alternative methods of estimating the impact of social programs: The case of manpower training, Journal of the American Statistical Association 84, 862-874.

Heckman, J., Ichimura, H., Smith, J. and Todd, P. (1998), Characterizing selection bias using experimental data, Econometrica 66, 1017-1098.

Krueger, A.B. and Lindahl, M. (1999), Education for growth in Sweden and the world, Swedish Economic Policy Review 6, 291-339.

Raaum, O. and Torp, H. (2002), Labour market training in Norway—effect on earnings, Labour Economics, forthcoming.

Regnér, H. (2002), A nonexperimental evaluation of training programs for the unemployed in Sweden, Labour Economics, forthcoming.

Smith, J. (2002), Special issue on program evaluation. Introduction, Labour Economics, forthcoming.