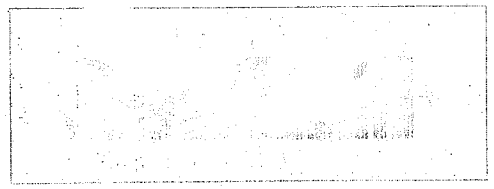
THE ECONOMIC AND SOCIAL RESEARCH INSTITUTE



THE STORY OF A SOCIAL EXPERIMENT AND SOME REFLECTIONS

ROBERT M. SOLOW

Thirteenth Geary Lecture, 1980

Copies of this paper may be obtained from The Economic and Social Research Institute (Limited Company No. 18269) Registered Office: 4 Burlington Road, Dublin 4, Ireland.

Price: £1.50

ISBN 0 7070 0039 4

Robert M. Solow is Institute Professor at the Department of Economics at Massachusetts Institute of Technology. This paper has been accepted for publication by ESRI, which is not responsible for either the content or the views expressed therein.

Printed in the Republic of Ireland by Mount Salus Press Limited, Dublin 4.

The Story of a Social Experiment and Some Reflections

It is a little known — deservedly little known — fact that I studied sociology and anthropology as an undergraduate before turning in despair to the Queen of the Social Sciences. I remember reading in those days about a figure who recurred often in the myths of many American Indian tribes. It was a god or demigod called "The Trickster". He would appear sometimes as a crow, sometimes as an eagle or a sparrow, sometimes as a wolf or otter or a fish and he would pester the poor Indians, causing bugs in computer programs, making the rivers run backwards, fiddling the order of nature, puzzling and confounding the Indians before vanishing as mysteriously as he had come.

Nowadays, most of my work is in macroeconomics, and I often feel as if "The Trickster" had decided to leave the Indians alone and do this thing to the macroeconomics profession instead: messing up the consumption function, introducing inexplicable glitches in the productivity trend, shifting the demand function for money just when you had come to rely on it. The worst consequence of "The Trickster's" machinations is that he pulls the rug from under the sober analysis. When economic behaviour is unstable, doctrine becomes unstable. There are usually two or more ways to explain the given set of erratic facts. The questions we want to ask are too complicated for the data to answer, given that "The Trickster" is at work.

One wishes that economics were an experimental science. The classical way to induce nature to part with the answer to a complicated question is to break the question down into simple

parts, and design a series of controlled experiments to explore the role of one factor at a time. The statistical theory of experimental design teaches us how to do a bit better than that, but the principle is the same. Unfortunately, that way out is closed to macroeconomics. All we have to go on is the one experimental run that history performs for us, and history never bothers to repeat itself holding constant all but one factor at a time. That being so, two clever macroeconomists can always find two models that will give equally good explanations of the narrow range of facts at our disposal, but have different implications for fiscal policy.

This line of thought gave me an idea for a Geary Lecture, when I had the honour of an invitation to give the 13th in a distinguished series. (I wonder if 13 is an unlucky number in Ireland too?) First of all, a plausible case can be made that Roy Geary *is* "The Trickster". He certainly has that characteristic habit of turning up sometimes as a coyote, sometimes as a salmon, now as a mathematical statistician, now as an applied statistician, once or twice as an economic theorist, several times as an analyst of social-accounting methods and concepts, and, more recently, as a student of wage differentials, unemployment, and the problems of the peripheral members of the labour force. I was especially interested to see how much of ESRI's recent and current research programme is aimed at this field of "social economics". That made me think I had a story worth telling.

I have recently been involved in a large-scale socio-economic experiment that has just come to an end after some four years. I would like to describe it to you both for its intrinsic interest and for its wider implications, which bear specifically on labour-market policy and, more generally, on social experimentation as a part of the policy process. Then at the very end, I will wonder out loud if this approach holds out any hope for macroeconomics.

One of the more intractable problems facing the US economy is the concentration of unemployment and low wages on a hard core of people who simply do not connect with the prime labour market. The men and women in question are usually residents of the decaying centres of large cities, and this fact is both cause and effect of urban decay; but rural poverty

persists too. They are often, but not always, young; they are usually uneducated. They are often, but not nearly always, black or Hispanic. Many combine two or three of these characteristics, and have a correspondingly harder time of it in the labour market. I do not suppose that the US is unique in having this problem. Indeed, it is my impression that migration within Europe and between Europe and its periphery has made Europeans familiar with the same complex of economic and social pathology. But the US has been diverse and geographically mobile for a longer time, and so we have been trying to do something about it for quite a while, not very successfully.

The generic name for the sorts of policies directed at this class of problems is Manpower Policy. We have had a long history of a variety of manpower policies. I would like to be able to tell you which of them had succeeded and which had failed, and what exactly it means in this field to succeed or to fail. That is not so easy to do, however, because most of the various schemes had been conceived in a hurry, translated into national programmes without much analysis or forethought, found disappointing in action even in the absence of clearly stated criteria, and abandoned either with a bang or a whimper, sometimes both. Worst of all, despite occasional attempts at evaluation, usually undertaken after the fact, the history of manpower policy has left behind it very little in the way of tested knowledge or reliable information about the operation of different programmes and their effects on the behaviour and labour-market experience of their participants.

My story has to do with a particular manpower programme that goes under the name "Supported Work"[1]. It began as a trial run conducted by the Vera Institute of Justice in New York in 1972. Vera's expertise is mainly legal, as its name suggests; but it is easy to imagine how it got involved in an attempt to provide employment experience for a group of ex-drug-addicts. The idea was to provide work experience as a

[1]For a more complete summary of the findings of this programme, see Board of Directors, Manpower Demonstration Research Corporation, *Summary and Findings of the National Supported Work Demonstration,* Ballinger Publishing Company, Cambridge, Mass. 1980.

bridge to the ordinary labour market. "Supported Work" had three distinctive features: the participants were organised to work in small teams consisting entirely of ex-addicts, so they did not initially have to cope with the problem of adapting to and being adapted to by the "straight" world; they were closely supervised, usually by someone who had been an addict, and "made it"; and the workplace demands made on the participants started light and were gradually intensified, so that absenteeism or lateness or malingering that might be excused at first would later on be cause for dismissal from the programme—this feature came to be called "graduated stress". After a limited time in the programme, participants were expected to "graduate" into the regular labour market. The emphasis was on the experience and habit of regular work rather than on training; no doubt some skills were acquired too, though most of the work performed was fairly low in the occupational hierarchy.

The early experience with supported work was favourable; and at this stage something innovative happened. A few perceptive people in the Ford Foundation and the US Department of Labour decided that the next step ought to be a large and carefully prepared experiment, with a formal research component, to study whether the supported work design could be generalised to other sponsors, other places and other client groups. Eventually the experiment was financed by the Ford Foundation and a consortium of half a dozen government agencies led by the Labour Department. It was designed, organised and managed by a non-profit corporation (Manpower Demonstration Research Corporation or MDRC) formed specifically for the purpose, although MDRC has since gone on to do other similar projects. The actual operation of supported-work enterprises was decentralised, as I will explain, and the actual research was contracted out, after a competition, to semi-academic firms specialising in that sort of thing. MDRC did, however, maintain close and active supervision of both field operations and the experimental design and subsequent research programme. I was, and am, a member, and vice-chairman, of the Board of Directors of MDRC for entirely serendipitous reasons. That is how I come to be telling you this story, so distant from my usual concerns.

Maybe "The Trickster" was ultimately responsible for this as for so many other things.

I could while away the time by telling you some details of the supported-work experiment. Like every other human enterprise it has accumulated a layer of anecdote and folklore. But I have to stick to the generalisable aspects of the process and its results; and for better or worse, Jersey City is not Cork, so the colourful details are irrelevant. There is, however, some organisational nitty-gritty that is worth mentioning, because I think it reflects the nature and limits of the process of social experimentation. In the end, the experiment was carried on at fifteen demonstration sites, of which twelve were in large cities and the others in wider, partly small-town and rural, areas. In each case the local enterprise was set up and managed by a local social-service organisation. Many of these organisations had been created, usually by an existing community agency, for the purpose of operating the supported-work enterprise; in some instances the existing agency operated the supported-work enterprise directly. The fifteen local entrepreneurs were selected in a competition from among perhaps three times as many applicants, the basis of choice being the merit of a proposed preliminary plan, and a reading of the managerial capacity available. This organisational feature is important, for the following reasons.

This tapping of local entrepreneurial talent is probably necessary in social experimentation, at least in the manpower field. Local conditions and attitudes vary, and the optimal size of each experimental site is probably small. It is hard to imagine successful day-to-day management by a centralised team of cloned bureaucrats. The result is an inevitable loss of *some* experimental uniformity. The lesson is that a deliberate effort is required to limit the loss of uniformity to what is tolerable without hopelessly compromising the value of the experiment. The problem of control is made even more difficult by the fact that the local entrepreneur sees himself or herself as being in the business of doing good, or perhaps the business of operating programmes, but in any case not as being in the business of producing statistical inferences. So the local operator is constantly tempted to do the Lord's work better by varying the experimental set-up; and the provision of finicky

detailed information for some highfalutin research bureaucracy comes low on the priority list. For us, however, reasonable uniformity and accurate information are the name of the game. We have found that the only practical way to maintain control over experimental conditions is to place the purse strings in the hands of the research organisation. There is nothing like being the chap who pays the piper if you want to be the chap who calls the tune. No doubt organisational habits and needs are different in different places — another application of the maxim that Jersey City is not Cork — but I think the tension I am pointing to here is quite general: you cannot run successful manpower programmes without tapping local initiative, but the scope you can allow to local initiative must be limited or the experiment, as experiment, will go down the drain.

There is an even more important generalisation that I want to emphasise, one that is absolutely central to the idea of social experimentation. Very early in the supported-work experiment, the advisory group that later became MDRC concluded that the experimental design would have to involve a formal control group. For example, any agency referring an ex-addict to the programme would be required to refer two at a time, and the intake process would randomly assign one of the two as an experimental and the other as a control. It was our intention to keep track of the members of the control group as best we could, to interview them periodically, and to pay them for the interviews in the hope of keeping in touch. The need for the control groups is related to the necessary decentralisation of the experiment. We want to find out if participation in supported-work improves the subsequent labour-market experience of participants. But, of course, the dominant influence on the later employment and wages of participants will be the later condition of the local labour market, as it is affected both by the business cycle and by conditions specific to Jersey City, or Atlanta, or Philadelphia, or Oakland. The only possible way to isolate the effects of supported work from these much louder background noises is to deal statistically with differences between the experimental and control groups who are exposed to identical extraneous conditions and differ only in their status with respect to

supported work. Attempts in connection with after-the-fact evaluation of other manpower programmes to construct artificial "comparison groups" have mostly failed, so the formal randomised control group seemed essential.

We were told it would never wash. We were told (a) that referral agencies and local operators would never stand for this cold-hearted exclusion of half the eligibles — we replied that the size of the intake was limited anyway, and all the available openings would be filled; (b) that it was immoral to deal thus with human beings — we replied that the size of the intake was limited anyway, and that the ultimate purpose of the experiment, to find out if supported work works, was in the long run best interest of the population at risk; and (c) that we would never succeed in tracking the control groups — we replied that we would sure as hell try, and in the event we succeeded well enough. Since we were interested primarily in participants' labour-market performance well after completion of the programme, interviews were scheduled at 9, 18, 27, and in some cases 36 months after enrolment, although the maximum stay in the programme was generally held to one year. Almost 70 per cent of the scheduled 36-month interviews were completed. There was attrition among both experimentals and controls, but not so much as to call the statistical results in question.

You will see, when I come to sketch the outcomes of the supported-work experiment, that we are trying to measure small and variable effects. Moreover, a fairly long series of follow-up interviews is essential, simply because it makes a big difference to one's judgement of a manpower programme whether its effects are ephemeral or enduring. Two consequences follow. The first has already been mentioned: there is no substitute for a rigorous statistical design, and this almost certainly means the creation of a formal control group, even against initial resistance. The second consequence is that sample sizes must be fairly large, first of all to permit some programme variations, but also because there is inevitably attrition in an essentially unstable population, and one needs to come to the last interviews with a sample size remaining that is adequate to measure the sort of effects one can reasonably expect to find. In our case, we made about 6,600 initial

random assignments, almost exactly half and half between experimental and control groups. Actually, as many as 10,000 people were employed as participants in supported work programmes, but many were not included in the research sample. There were several sources of this discrepancy: in some cases programme operators had good reasons for making programme variations that simply did not fit into the experimental design, so the corresponding participants had to be excluded from the research sample; and at 5 of the 15 sites, MDRC concluded that valid data could not be expected, so we continued the sites as a sort of demonstration exercise, if only because a commitment had already been made, but again excluded those sites from the experimental design.

With such numbers, I hardly need tell you that experimental social research is an expensive business. The supported-work experiment was in the field for just under four years, with the research work continuing for at least another year. During that time, site operations spent some $66 million, with research and administrative costs bringing the total up to about $80 million. About a sixth of the site expenditures were covered by the sale of goods and services produced by supported-work enterprises. About a third of the dollars spent by the operating sites were raised by them from locally-available social welfare funds; when I called the local operators "entrepreneurs", I meant it. The remaining half of site expenditures plus all of the research and central management cost was borne by the original sponsoring consortium. That is expensive knowledge, but I am convinced the price is worth paying. Governments must be made to realise that even such large sums pale into triviality next to the much larger amounts that get poured down rat-holes in the belief that it is more important, or at least better politics, to respond hastily and visibly to social needs on the basis of no tested knowledge at all.

It is time I told you what we actually found out. I have already mentioned that the apparently successful prototype of supported work was aimed at ex-drug-addicts in New York. Our goal was to see if the programme would work in other places, with alternative client groups. Altogether 15 sites operated all around the country, of which 10 were full-dress research sites. We included four client groups this time. There

were ex-addicts at 4 sites. Ex-convicts (ex-offenders is the standard euphemism) were enrolled at 7 sites. Seven sites enrolled groups of AFDC mothers. (Aid to Families with Dependent Children is the largest "welfare" programme of the Federal government. The AFDC group consisted of women who were currently receiving money under this programme and had done so for at least 30 of the preceding 36 months; women whose youngest child was under 6 years were excluded.) Finally, 5 sites enrolled groups of youths between 17 and 20 years, without a secondary school degree or equivalent, out of school for at least the past 6 months, and with some record of delinquency. Obviously, most of the sites accepted more than one of the client groups.

The intake data showed that the experiment's intention, to try out supported work on groups whose *a priori* prospects in the labour market were very bad indeed, was met with something to spare. If ever people could be described as "disadvantaged", these could. I won't bore you with the details, but well over three-quarters were black or Hispanic, and fewer than half in each target group had worked at all in the year preceding enrolment.

There is much to be said about the experience of our sample while they were enrolled in the programme. This would be relevant if supported work were thought to be, potentially, a permanent or long-term way of life for these or other sub-groups of the population. From this point of view, supported work would be analogous to the sheltered workshops available to the physically-handicapped, or blind, or retarded, in some countries of Europe and occasionally in the US. It is true that the impetus for the experiment arose from the possibility that a period of supported work would ease the transition into the "straight" labour force for members of the client groups. But the performance of supported workers while in the programme is of some interest, if only for its bearing on the cost of the programme; so I shall say a word about that before turning to the longer-run post-programme results.

The various work sites generally tried to sell their services by some contract arrangement, though not usually at a competitive market price. By the end of the demonstration, three-quarters of work days were generating at least some

revenue. Half of all the effort went into a variety of clerical, building maintenance, business and other services. (There is something entertaining in the picture of a crew of ex-addicts painting a police station.) A little over a quarter of project days went into construction-related activities, including the rehabilitation of run-down buildings. Two of the sites developed successful manufacturing enterprises, but overall only eight per cent of project days fell into the manufacturing category. These differences reflect management style and local opportunity rather than anything more fundamental.

Trying to value the output of supported workers for benefit-cost calculations is no easy matter. Little of it was sold competitively, and of course much of the work fell short of standard quality. We tried several methods of evaluation, including the use of knowledgeable building tradesmen and others to assign a market value to work performed. Depending on client group, the value of in-programme output per participant ranged from $3,000 (ex-offenders) to $4,500 (AFDC mothers). In each case, this figure fell a few hundred dollars short of local supported-work costs (materials, supervision and local overhead). So the wages of supported workers were, approximately, a pure transfer. This does not strike me as forbiddingly costly; but clearly any major net benefit from the programme must come from indirect benefits and, more important, enduring post-programme effects. I turn now to those.

I have to describe the post-programme effects separately by client group because they differ and differ substantially. That is a pain in the neck, but an important general conclusion. It may have been known, or at least vaguely intuited, by experts. But it was news to me, and an important lesson for legislators, so it is useful to have evidence. There is probably no such thing as a generally effective manpower programme in a diverse society. They have to be tailored to their clients and probably to local social and economic conditions as well.

We pursued our experimental design for 27 months after intake, and in some cases 36 months, although time in the programme was limited to 12 months, and averaged less. The AFDC group came out best. The women who participated in the programme performed significantly better than the

controls in terms of increased employment, increased earnings, and reduced dependence on public assistance. These differentials remained statistically significant throughout the 27-month period of observation. One's impression is that the experimental-control differences in per cent employed, hours worked, and monthly earnings had more or less stabilised after 27 months. The overall cost-benefit calculation, including a necessarily shaky allowance for the discounted value of future differences, was clearly favourable. For this group, the indicated conclusion has to be that supported work works. I am not sure we would have guessed that at the outset.

The conclusions for the ex-addict group are more complicated, but just as interesting. (Remember that the Vera prototype was confined to this client group.) The experimental-control differentials in percentage employed, hours of work, and total earnings dwindled away to nothing by 24 months after initial intake. But then they turned favourable again, smoothly enough so that one is inclined to believe that something systematic is happening. In the sample interviewed 30-36 months after enrolment, all three differences (employment, hours and earnings) are statistically signficant at the five per cent level, and the mean differences are not at all trivial. It is obvious that a longer follow-up is needed to confirm this trend and to understand it. At a minimum it is promising. We found also a substantial reduction in criminal activity in this group as compared with controls, during their time in the programme and afterwards. Only some of these differences are statistically significant, but the pattern is so consistent that it is hard not to believe in their reality. (By the way, it is easy to generate healthy scepticism about self-reported criminal activity, even with a pledge of confidentiality. We put some effort into a cross-check with police records and found that there was indeed under-reporting of criminal activity, but with no significant difference in that regard between experimentals and controls. So the self-reported differences may have high variance with little or no bias.) When social benefits from the estimated reduction in criminal activity are included, the benefit-cost calculation for ex-addicts comes out strongly favourable. The reduced-crime benefits are very substantial, and evidently call for further research.

The next sub-group consisted of ex-offenders. Each eligible enrollee had been in prison within the six months preceding enrolment, as result of a conviction: 95 per cent were males, 90 per cent were black or Hispanic, 11 per cent reported that they had never worked — average age was 25 — and half had not worked at any full-time job during the past six months. Average earnings during the preceding year were $580, the result of less than six weeks of work. The average number of arrests per participant exceeded nine, with previous time in jail averaging almost 200 weeks. These ex-offenders tended to drop out of the programme more often than others; the average stay was 5.2 months. Although participants had a somewhat better earnings and employment record than controls after 27 months, the difference was not statistically significant. Unlike the ex-addicts, the ex-offenders who participated in the programme showed no reduction in criminal behaviour. When everything is added in, we come to a net-benefit total that centres on zero, with rather a wide range of possibility on either side. For ex-offenders, supported work cannot be regarded as a success.

Our last client group was limited to young people, 17-20 years old, who had dropped out of school. We insisted that at least half have a record of delinquency or crime. Over a fifth of the sample had never worked, most were males and only a small fraction were white. So far as our data go, supported work had no significant long-term effect on the employment, earnings, criminal activity or drug use of the youth group. Whatever the problem here, supported work is not the solution.

For the practical-minded person interested in social policy, these results contain a message. Supported work is an extraordinarily promising device for the integration of welfare mothers into the labour force. It is clearly worth trying with ex-addicts; and any real-life trial should contain a strong data-gathering component to check on the experiment's indications of favourable in-programme and post-programme effects on criminal activity, and apparent, but uncertain, long-term improvement in employment and earnings. For ex-offenders, there is no sound basis for rejecting the null hypothesis that the programme has no effects; one might perhaps advocate further

experimentation, especially with an improved programme tailored especially to the needs of this client group, but with no implied promise of eventual success. In the case of the youth group, one must conclude that supported work offers no special advantages.

I think it is clear that social policy in this field can now proceed on an incomparably firmer basis than would have been available without the supported-work experiment. If there is any merit to rationality in manpower policy, then we have shown that the carefully designed social experiment is one way to achieve it. So far, I have not heard of any other. I want to emphasise that the conclusions I have quoted could not have been intuited or predicted in advance. Experts in the field had no way of knowing if the scheme would work at all, and certainly no basis for distinguishing among the various client groups. This was not a matter of demonstrating the obvious. By the way, there is room for further experimentation with other client groups. Supported-work enterprises are now under way employing ex-alcoholics and mentally retarded clients; they are not part of an explicit experimental design, although they do collect internal data that might conceivably be useful.

I want to emphasise that the effects measured in this experiment tend to be small in absolute terms, even when they are statistically significant. Neither supported work, nor anything else that I have heard about, provides the "dramatic relief" so prized by the manufacturers of pills. Needless to say, I am sceptical about the claims made for pills. But I am just as concerned about the need for truth in advertising when it comes to employment policy. I do not know how it is in Ireland, but in the United States the political process usually follows a predictable and unproductive sequence. To generate any action at all, a failure in the labour market has to be overblown into a "crisis". Naturally, a crisis calls for immediate action, and immediate action necessarily implies the legislation of an untested programme. To make an untested programme sound like a worthwhile solution to a crisis, inflated claims have to be made for its effectiveness. In due course, the programme fails to live up to those claims. The likeliest outcome of a "crash programme" is a crash. No doubt a certain amount of good gets done by this process. But the ever-

present danger is that the history of unfulfilled promises ends up by discrediting the whole idea of social policy. That appears to be happening in the United States now.

In the case of supported work, as I said, the measured effects are generally small. Only in the case of the AFDC client group can we state with some confidence that the net benefits of the programme are positive. The next most favourable case, that of the ex-addicts, depends heavily on some intrinsically uncertain findings about reduced criminal activity. When it comes to post-programme performance in the labour market, the statistically significant effects are hardly dramatic, although they appear to be real and non-trivial. Let me quote just a few examples: 19-27 months after enrolment 49.1 of the AFDC mothers in the experimental group were employed, as against 40.6 per cent of the controls. The difference of 8.5 per cent is significant at the 5 per cent level. The difference in average monthly earnings was $77, $243 as against $166. Amongst the ex-addicts the 19-27 month interviews showed essentially no difference in per cent employed, hours worked, or monthly earnings. In the 28-36 month interviews, 64 per cent of the experimentals were employed versus 54 per cent of the controls, with average monthly earnings of $326 versus $224. The employment difference is significant at the 10 per cent level, the earnings differential at the five per cent level. Among the ex-offenders, the 28-36 month interviews showed a differential in average monthly earnings of some $60 favouring the experimentals, but that difference fails of significance at the 10 per cent level.

No one who cares seriously about the employment of disadvantaged groups will sneer at those results. They are not at all trivial. I quote them in order to make two points. First, effects of that order of magnitude can *only* be won by carefully designed experiments with substantial sample sizes. They are simply too small to be detected by casual observation. Second, somehow we have to learn to make a convincing case for policy initiatives based on reasonable estimates of the probability of success and the quantitative meaning of success. If the professional policy community allows itself to promise more than it can deliver, it will end up not delivering what it promises, and eventually the promises will be disbelieved and there will be no delivery at all.

I could tell you more about the details of supported work and the measured outcomes of the experiment. But I could not tell you enough to make that a worthwhile enterprise. Besides, this experiment, like any worthwhile social-science experiment has been well-documented and thoroughly reported. For anyone who really wants to know the details, the published reports are the logical source. For my part, I want to draw two sets of general conclusions.

The first set concerns the process of social experimentation itself. The example I have described to you had as its object the evaluation of a policy device. One could equally well imagine experiments whose object was the pursuit of knowledge about social behaviour, without any explicit application to policy directly at stake. My guess is that most serious experiments will be policy-directed, perhaps with rare exceptions. The reason is simply that serious experiments are very expensive, as I have already pointed out and intend to point out again. That means, in turn, that all or most serious social experiments will have to be financed by central governments; and they are unlikely to spend large sums on the pursuit of small bits of social-scientific knowledge for their own sake, at least not until the method has proved itself again and again. But the experimental evaluation of social policy devices, in advance of full-scale operation, is a very valuable process and should be used very much more often than is the case now. Of course, it will often prove possible to piggy-back some questions of intrinsic scientific interest on an experiment that is primarily policy-oriented. For example, MDRC thought for a while about introducing some systematic wage-variation into the supported-work experiment. We decided against it, because we feared that any extra noise might mask the main effects. So all supported workers were paid a starting wage close to the statutory minimum, subject only to merit increases, as in the straight labour market. It would be useful for social scientists to get used to thinking in those terms.

Once an experimental approach to policy evaluation is in the cards, the supported-work experience suggests some guidelines for experimenters. First of all, as I have tried to emphasise, careful experimental design is absolutely vital. There is room for "demonstrations", whose purpose is only to

see if something can be done at all. But as soon as inference about effects, or about costs and benefits, enters the picture, the general presumption should be that the effects may be small, and are likely to be buried in the noise unless care is taken to design an experiment that will isolate them. Secondly, social experiments have to be on a "real" scale — by which remark I mean to say that playing for pennies will not reveal true behaviour responses. Toy experiments will not stimulate the real world. Third, since it may be necessary to detect rather small effects, large samples will generally be required. Sample sizes should be determined as part of the experimental design, and they will depend in the standard statistical way on the number of questions to be answered, the minimum effect one wants to be able to detect, and the operating characteristics — significance level and power — one is prepared to settle for. Even then, I believe in safety margins. The temptation to settle for smaller samples is always present — beggars can't be choosers — but it is to be resisted, or the whole concept will be discredited. Fourth, for most social-policy evaluations, long-term effects are important, and so experiments have to allow for long-term follow-up of experimental subjects and controls. This will add even more expense to what is already an expensive process, but it is worth the trouble. Fifth, I think it is important that the research component of any social experiment be designed into it from the beginning, not tacked on after the rest of the decisions have already been made. I have even suggested that the researchers control the enterprise, even to the extent of doling out the money. In any worthwhile social-policy experiment, the research component will always appear to the operators and participants as a useless appendage that merely gets in the way of the real business at hand. That may be true of social policy; but in social-policy experiments, the research is paramount.

There is one last point I should make, and I will make it separately to emphasise its importance. Social experiments are experiments involving people; the subjects are not likely to be injured or made sick or endangered. But, nevertheless, it is absolutely vital that there be clear and explicit standards safeguarding the safety and integrity of the people involved. They should not be humiliated or tricked or used against their own

interests. And there should be a powerful mechanism for verifying and enforcing the standards.

Now, finally, I would like to turn briefly to another kind of question. I am, after all, an economic theorist by trade, and primarily a macroeconomic theorist at that. As if macroeconomics were not difficult enough, in recent years some of the ordinarily reliable empirical relationships have begun to turn erratic. Moreover, and even more troubling, it has been increasingly the case that when any important macroeconomic question is posed econometrically the brotherhood is able to come up with two answers, namely, "Yes" and "No". My own feeling is that the advance of economic theory has led us to ask more and more subtle questions of the data, usually of data in the form of fairly short-time series. It is likely that we have outrun the capacity of the data to provide answers. These days, when two econometricians are able to provide mutually contradictory answers to an interesting question, rather than asking myself who is right, I am increasingly inclined to conclude that the data simply do not speak intelligibly.

There are then two possible ways out. One is to appeal from short-time series to long-time series. In some cases that means using less satisfactory historical data; in others it may mean waiting for history to produce additional observations. The trouble with that resolution is not only that waiting is hardly a solution; one may also suspect that long-term series, even when they exist, may be irrelevant or worse, because the underlying parameters and relationships may not be stationary, may indeed be more likely to change the longer the time interval.

The second way out is to look for one's basic knowledge in micro-data. Is it possible that the social experiment may be the answer to the macroeconomist's prayer? To take a current, if slightly far-fetched example, if the aggregate demand for money proves unstable, should we be trying to estimate its parameters from microeconomic, even perhaps experimentally obtained, data? I think that this is an illusory hope. Experience suggests that the degree of inter-individual variation in micro-data is extraordinarily large. The results of the supported-work experiment are an excellent case in point. Even with a large sample the main effect is almost buried in the noise. Everyone who deals with cross-section data is used to their low resolving

power. I suppose a determinist would say that it happens only because we are unable to control for the many exogenous variables affecting any individual's response to a change in interest rates or in anything else. Saving and hoarding habits may, for all I know, go back to childhood experiences, maybe even to toilet-training. Mothers have been blamed for everything else, why not for peculiarities in the demand for money? It doesn't matter. What we cannot observe and control for is to all intents and purposes noise. I have tried to indicate the power of micro-experiments for certain kinds of policy studies. But it is not my impression that macroeconomists can find salvation in estimating relationships from micro-data and aggregating afterwards. We shall just have to learn to live with "The Trickster".