JOURNAL OF APPLIED ECONOMETRICS, VOL. 12, 459–465 (1997)

DESIGN OF THE EXPERIMENT

JAN R. MAGNUS[a] AND MARY S. MORGAN[b]*

[a] *CentER for Economic Research, Tilburg University, PO Box 90153, 5000 LE Tilburg, The Netherlands*
[b] *University of Amsterdam and London School of Economics, Houghton Street, London, WC2A 2AE, UK*

SUMMARY

In this paper, we give a short history of our field trial experiment in applied econometrics and outline the aims of the experiment. We discuss features of our experimental design, and judge to what extent our design was successful. The difficulties of such an experiment are acknowledged. The issue of tacit knowledge in applied econometrics is raised as a problem for further research. © 1997 John Wiley & Sons, Ltd.

J. Appl. Econ., **12**, 459–465 (1997)

No. of Figures: 0. No. of Tables: 0. No. of References: 9.

1. THE QUESTION BEHIND THE EXPERIMENT: A SHORT BACKGROUND HISTORY

For the last ten to fifteen years, econometricians have been debating amongst themselves the right way to do econometrics. The discussions have ranged widely over questions about the concept of probability appropriate for economics, associated statistical inference procedures, criteria of economic versus statistical significance, the kinds of models appropriate for econometrics, the process of modelling, and so forth. The debate has been joined at many levels, from methodological discussions to arguments over the rival merits of software packages.[1] The debates have been informative, and in marked contrast to other areas of methodological debate inside economics, they have often managed to combine both philosophical sophistication and practicality (see, for example, the papers on econometrics contained in Hamminga and De Marchi, 1994). These discussions are, without doubt, a healthy sign of an active econometrics profession. Apart from the critiques, the arguments have been connected to some thoroughly innovative developments in econometrics: for example, methods to compare models, to compare forecasts, to assess the usefulness of calibration versus estimation, and so forth.

The problem is not a new one. Indeed, there have been periodic debates within economics about how to do econometrics, going back to the Tinbergen–Keynes debate of 1939–40 and the 'measurement without theory' debate of 1946–7 (see Hendry and Morgan, 1995). But the current debate has gone on longer and been more fruitful than earlier ones in generating a greater number

* Correspondence to: Mary S. Morgan, Department of Economic History, London School of Economics, Houghton Street, London WC2A 2AE, UK

[1] The debates have been so endemic at so many levels that it is difficult to give a concise set of references. The classic critical survey of three of the main approaches is Pagan (1987).

CCC 0883–7252/97/050459–07$17·50 *Received 11 July 1997*
© 1997 John Wiley & Sons, Ltd.

of clearly established methodological positions and practical methods and tools associated with them. Indeed we see in these new developments real attempts to make tools and methods in line with methodological pronouncements. These have been conveniently enshrined in software packages which enable econometricians to combine the levels of econometric theory, methodology and data treatment in one approach.

Despite the activity of debate, it has been somewhat disquieting that few matters have been resolved. Nevertheless, econometrics remains a very important part of applied economics and plays a vital role in economics. Because of this, the establishment of a credible applied econometrics is possibly the most important task of econometrics today.

Discussions about how to do econometrics have been dominated by the view that there are 'right' and 'wrong' ways to do things, that the matter is one of agreeing on *the* correct methodology rather than of assessing the practical usefulness of different approaches in applied circumstances. Our aim has been different (and is much in line with the recent more practical bent developing in the debate), namely to assess different ways of doing econometrics by a controlled field trial experiment in applied econometrics. The basic idea of the experiment is very simple, namely to take a specified data set and let several researchers carry out the same set of applied econometrics tasks but with different methods, approaches and beliefs. Our overall aim is to assess, within the environment of our experiment, the differences between the several ways of doing econometrics in a practical application.

This sounds simple enough, but the design of such an experiment is not simple; indeed, our experience suggests that it is fraught with difficulties (on which more below). We have learnt much from conducting this experiment, which is to say that there are many aspects of our design that could be improved. No doubt we made many beginners' mistakes, the most important one of which was that we were too ambitious: we tried to do too much in one experiment. This was partly due to our lack of experience, but also because of the history of our attempts to get this experiment off the ground. This had proved so difficult that when we finally succeeded we were tempted to try to do as much as possible with our opportunity.

As far as we know, this is the first time such an applied econometrics field trial experiment has been undertaken. Ed Leamer planned an experiment in applied macro econometrics (to take place in 1990/1), but this never took place. Jan Magnus first proposed a field trial experiment for econometric methodology in 1991. But despite being highly rated as innovative, appropriate and important by the Netherlands Organization for the Advancement of Research (NWO), the project was not funded. A similar fate met further attempts by the two organizers (Magnus and Morgan) to obtain funds from the Economic and Social Research Council (ESRC) in the UK in 1992–3 and 1993–4. Assessors agreed it was a good idea and that the experiment was sufficiently well thought out to be worth supporting, but we failed to get the funds necessary. We were delighted when the opportunity to do part of the planned experiment was made possible by the support of the editorial board of the *Journal of Applied Econometrics*, and are grateful for the opportunity this has given us to conduct the experiment.

We are also grateful to CentER for their funding of the workshop where the experimental results were discussed. We have been impressed by the commitment of those who participated in the experiment — those who entered the field trial, reported their results, and revised their reports to meet our exacting deadlines. We have also benefited greatly from the help of our band of assessors who worked hard to think about how the experimenters obtained different results. All those involved genuinely entered into the spirit of the experiment, and the success of the field trial, if we can call it successful, is largely due to their combined commitment.

 © 1997 John Wiley & Sons, Ltd.

2. THE EXPERIMENTAL DESIGN

The aim of the experiment was to assess competing methodologies of econometrics by an applied field trial. Such a field trial posed a number of design problems, for the main problem of field trials is to establish sufficient control in the design that one can, in principle, expect to learn from the results.

In the classic cases where field experiment design techniques have been worked out (see Fisher, 1960 (originally, 1935)) there is always an untreated group which acts as the control against which the difference in outcome can be measured. In our case, there appeared to be no possible 'control group' or 'untreated group' against which to measure the differences of applying one econometric methodology rather than another. In order to establish a neutral base line against which to judge the different approaches we picked a 'classic paper' in applied econometrics, one which everyone could agree had been, at its time, the best applied econometrics could offer. The chosen paper was James Tobin's 'A statistical demand function for food in the USA', published in 1950 in the *Journal of the Royal Statistical Society*, Series A. This solution, and our associated choice of classic paper, had the immediate implication that we widened our aims. Our aims, as we expressed it in our experimental information pack were to assess not only 'the differences between the several ways of doing econometrics in a practical application' but also, since Tobin's paper was not a contempory one, 'the advantage (if any) of 45 years of econometric theory since Tobin's paper' and 'the impact of new economic theories' (Magnus and Morgan, 1995b, p. 3).

Another level of design control required us to hold constant as many as possible of the other circumstances which might vary, or at least hold them equal between participants, so that one can see more clearly the differences due to different methodologies. There are three main areas of experimental control we tried to instil in the design here.

First, we constrained participants to use only the data we provided. We aimed to control for particular local knowledge by supplying data from two economies: the USA and the Netherlands, believing that no participant would have expert local knowledge of both.

Second, we had to choose the qualities or criteria on which the methodologies were to be assessed; that is: What should a practical econometrician be able to do in working on this data set? Here we believed that there were a variety of things: propose the most suitable model, forecast, produce estimates, investigate policy implications, etc. This required that we set a well-specified set of tasks so that each methodology could be assessed as far as possible on the same tasks.

Third, we feared 'contamination' between our participating 'subjects', and so asked them not to communicate with each other, or publicly report the results, until after a specific date by which all experimental reports were due.

Our experimental design was complicated by the fact that, although we wanted comparability in order to be able to assess the methodologies on the same tasks, we were quite aware that the existing methodological approaches were, in part, designed for different kinds of problems and different types of data sets. We thus had to satisfy two objectives in our design: comparability and diversity, and these were relevant both for our choice of task and of data sets.

On the one hand, we wanted the data to be sufficiently simple so that we could isolate the impact of the changes in technology and theory since 1950. On the other, we wanted sufficient richness in the data sets so that the study was interesting for its own sake (to motivate participation) and would allow the participants to shine in their own specialities. We therefore chose, in

© 1997 John Wiley & Sons, Ltd.

addition to Tobin's data set, two further sets of data. The data used in the experiment thus consisted of:

(a) The set used by Tobin
(b) Further data (after 1950) for the USA
(c) Data of the same type, but from the Dutch economy.

The US data consisted of both time-series and cross-section data in the same spirit as Tobin's data. This data set was somewhat narrow, but it fulfilled our requirement to enhance comparability of the various approaches. The Dutch data set, on the other hand, was much richer. This set combines time series (1948–88) data with three budget studies, two of which are quite detailed. (A brief description of the data is included in this issue. Full details are given in Magnus and Morgan (1995b) or in Magnus and Morgan (1998, forthcoming).) Taken together, the US and Dutch data sets involved time series, cross-sections and household level data and thus satisfied our requirement for diversity in the raw material.

The tasks were set to allow for the full range of what one might expect practical econometrics to achieve, and to allow different methodologies, even specialized ones, to show their best efforts. The tasks are fully described in our 'Organization of the experiment' (this issue). Suffice it to say here that they involved a measurement task using the US data directly comparable to the one that Tobin had undertaken; a 'measurement with related information' task involving the Dutch and American data; a forecasting task involving the Dutch data; a policy question task, and a self-chosen task. Since we were interested in assessing the advantages and disadvantages of the different methodologies, there was, of course, freedom on how the tasks were performed. And because of the specialisms associated with different methodologies and with participants' own skills and interests, we did not expect all participants to undertake all tasks.

In summary, we selected Tobin's 1950 paper as an example of 'good applied econometrics' to provide the baseline for the participants' own analysis. The experiment was designed to show first what difference modern approaches and technologies make to Tobin's results. Beyond that base level, we wanted to generate information which would enable comparative assessment of the different approaches. By constraining the participants to the same data set and asking all participants to carry out a well-specified set of 'tasks' (how they performed the tasks was up to them), the experiment aimed to provide the raw material for explicit and constructive assessment.

3. TACIT KNOWLEDGE

It had been an essential part of our earlier attempts to obtain funding for the experiment that we would also use the opportunity to make a first attempt to assess the role of tacit knowledge in econometrics. Such practical knowledge fills the gap between methodological treatises and successful applications. This difference between theory and practice has led some unsympathetic observers to claim that 'econometricians do not practise what they preach'! This, of course, is naive. Applied econometrics, like all applied science, involves craft skill, or tacit knowledge, which enables the master econometrician to get practical results from using his or her methodology where the unskilled worker flounders. Although we were unable to build in any substantial study of tacit knowledge in this experiment, we hoped to make a first pass at this issue by asking all our participants to keep logbooks of the process of how they did their work. This is discussed in 'Organization of the experiment' (this issue). (A separate, but related, experiment on tacit

 © 1997 John Wiley & Sons, Ltd.

knowledge will be reported in the book-length report on this experiment—see Siegert and Magnus, 1998, forthcoming).)

4. BRIEF OUTLINE OF THE FIELD TRIAL ARRANGEMENTS

In the May 1995 issue 1995 of the *Journal of Applied Econometrics* we announced and briefly described the experiment (Magnus and Morgan 1995a). A considerable number of participating teams (groups or individuals) 'registered' and received, in July 1995, an experiment pack giving full information on the experiment (tasks, rules, data description, etc.), a data diskette and a reprint of the Tobin article. Eight experimental reports were received back in the summer of 1996. In December 1996, experiment participants, assessors and the organizers met for a workshop at CentER for Economic Research, Tilburg University, to discuss the individual reports and results of the experiment. These findings are reported, in part, in this special issue. Six of the eight papers (together with comments of the assessors and replies by the authors) are published in this special issue. All eight reports and more extensive comments and analysis will appear in Magnus and Morgan (1998, forthcoming).

Full details of the process of the experiment—rules, tasks, conduct and assessment—are included in 'Organization of the experiment' (this issue).

5. ASSESSMENT OF THE EXPERIMENT

From the beginning of the experiment we planned to ask the help of a panel of independent experts for commentary on the experiment and to help us to assess the reports from our participants in the field trial. Of course, we tried to get the best assessors and to minimize bias. In order to avoid the problem of having assessors who were all in one 'camp', and to make sure each participant had the chance of nominating assessors who were sympathetic to their approach, we asked each participant to nominate two assessors. (In fact, few participants nominated assessors, and not all assessors nominated by participants, or selected by ourselves, were willing to participate.) As a further procedure to minimize bias, we planned to ask each assessor to comment on several papers, not only on the one for which they had been nominated. We intended to ask the assessors to act as commentators on the submitted reports and as referees for publication purposes. In the event we found that our assessors were most responsive and helped us all to think about why results from the different teams differed by making comparative comments on the reports they were discussing.

By asking a panel of independent assessors to comment on the reports, we hoped this would maintain our own position as neutral between participants in the field trial. We believed it would be invidious for us to be assessing the individual reports as well as running the experiment. We were also aware that the *Journal of Applied Econometrics* provided an important, professionally neutral, place to report the results, but again, only if we ensured its neutrality with respect to the different approaches of the participants in carrying out our editorial responsibilities. We therefore announced that, as guest editors of the special issue, we would be assuming complete editorial responsibility and that our aim for the special issue was that it would contain the most interesting reports, as judged by the independent assessors (acting as referees). In addition, of course, all reports for publication had to meet the academic standards and procedures normal for submissions to the *Journal of Applied Econometrics.*

© 1997 John Wiley & Sons, Ltd.

6. HOW SUCCESSFUL WAS IT?

There were three major areas of the experiment which in retrospect we ought to have thought about more carefully beforehand. One was that we asked the experimenters to attempt too many tasks. We surmise that this was one of the reasons why a large number of signed up participants did not take part. More seriously, it meant that in the experimental reports so many different things were done that legitimate areas for comparison became more limited. This is, as we have already noted, a classic problem of the beginner — being over-optimistic and over-ambitious about what can be learnt from one experiment.

In addition, our ability to build in controls proved insufficient. Recall from our experimental design discussion (above), that the aspects we aimed to control were the data used, the tasks set, and to prevent both cross-contamination of knowledge between experimenters and local knowledge advantage. In practice, experimenters (unconsciously) circumvented these controls, for example by working on different time periods from the time-series data. The tasks were not as clearly specified as we believed in advance, so interpretations led to results which were not always (at least at first sight) clearly comparable. (And one of our tasks, the policy task, was found by all concerned to be thoroughly misconceived!) Moreover, we had not allowed for (or even considered) the possibility, pointed out to us by one participating group, that experimenters would seek to differentiate their own contribution from others sharing the same methodology who were among those reported to be taking part (the experimental pack contained a list of all those who had originally committed themselves to the experiment). Specific knowledge was clearly at work, but it was not clear that there was local advantage: we found that one of our Dutch teams knew more about the economic history underlying the American time-series data than did one of the American teams.

Even with the simplest first task of measurement, we had underestimated the ability of eight participating teams to produce different versions of the variables, different models, and different measurement procedures. The difference in 'methodological approaches' that we were trying to get at included a mix of econometric tools, methodologies, specialization of interest and economic beliefs and theories. This was a potent mix which led to a huge variability of outcomes and made our tasks and that of the assessors much more difficult.

The third failure was not to foresee that this variability in outcomes was the necessary outcome of the breadth of uncontrolled variability inherent in these 'methodological approaches'. One reason that we did not foresee the problem clearly was that we had not adequately worked out beforehand how we were going to assess the experiment. We were aware, and had taken account of the problem, that at the technical end, different methodologies required different methods of assessment (and indeed this was one reason we needed a team of assessors with different expertise). But we had not fully thought through the broader problem of comparison of the results. We thus broke one of the cardinal rules of experimental design — that of working out what kinds of results might be obtained with our experimental set-up, and then refining the design until it was such that we ought, in principle, be able to draw valid inferences about the subject of our experiment. Maybe such an ideal experiment was never practically possible.

On the plus side, we count it as a success that the reports from the experimenters do not follow the normal style of applied econometrics papers — reporting only finally successful results, and giving little idea of the process. It is not only the presence of the logbooks which adds verisimilitude to the accounts, the experimental reports in this issue are just that: they have a degree of immediacy and transparency which is not to be found in journal papers.

 © 1997 John Wiley & Sons, Ltd.

Despite the failures of the experimental design, the setting up and running of the experiment was an exciting and, we believe, worthwhile experience. A lot has been learnt about the problems of such experiments. It remains to discuss the comparative results of the methodologies involved in the field trial. Initial comments by our assessors are found in the pages that follow each report. Much more explicit comparisons and assessments will have to await the book-length report on the project in Magnus and Morgan (1998, forthcoming).

REFERENCES

Fisher, R. A. (1960), *Design of Experiments*, 8th edition, Oliver and Boyd, Edinburgh and London.

Hamminga, B. and N. De Marchi (1994), *Idealization in Economics*, Rodopi, Amsterdam.

Hendry, D. F. and M. S. Morgan (1995), *The Foundations of Econometric Analysis*, Cambridge University Press, Cambridge.

Magnus, J. R. and M. S. Morgan (1995a), 'An experiment in applied econometrics: call for participants', *Journal of Applied Econometrics*, **10**, 213–16.

Magnus, J. R. and M. S. Morgan (1995b), *The Experiment in Applied Econometrics: Information Pack* (54 pages), CentER for Economic Research, Tilburg University.

Magnus, J. R. and M. S. Morgan (eds) (1998), *The Experiment in Applied Econometrics*, John Wiley, Chichester and New York, to appear.

Pagan, A. (1987) 'Three econometric methodologies: a critical appraisal', *Journal of Economic Surveys*, **1**, 3–24.

Siegert, W. K. and J. R. Magnus (1998), 'Tacit knowledge in econometrics', in J. R. Magnus and M. S. Morgan (eds) (1998), The Experiment in Applied Econometrics, John Wiley, Chichester and New York, to appear.

Tobin, J. (1950), 'A statistical demand function for food in the U.S.A.', *Journal of the Royal Statistical Society*, Series A, **113**, Part II, 113–41.

© 1997 John Wiley & Sons, Ltd.