10 Evidence-based policy: summon the randomistas?

Andrew Leigh[1]
Research School of Economics, Australian National University

Abstract

Randomised experiments are a standard toolkit for the evaluation of new medical treatments, but are underutilised in Australian policy evaluation. Drawing on examples from Australia and overseas, I discuss the strengths, limits, ethics and politics of randomised trials. In a federation, it may be effective for the national government to establish a central fund for states and territories to conduct randomised policy trials.

10.1 Let's start with a vitamin tablet

For the past 10 years, I have taken a multivitamin pill once a day with my morning coffee. Like any good morning habit, it is a comfortable and familiar routine. But I've gotten a warm glow to know that each day begins with a simple act that helps take care of my body.

Earlier this year, a friend suggested that I read an article published in the *Journal of the American Medical Association* (Bjelakovic et al. 2007).[2] The authors set out to answer the question: do vitamin supplements make you live longer? To answer this,

1 Email: andrew.leigh@anu.edu.au. Web: http://andrewleigh.org Parts of this article draw on a keynote lecture delivered at the NSW Bureau of Crime Statistics and Research 40th Anniversary Symposium on 19 February 2009. I am grateful to participants at the Productivity Commission Roundtable, Terry O'Brien, Angela O'Brien-Malone and Nikki Rogers for valuable comments on earlier drafts. Jenny Chesters and Susanne Schmidt provided outstanding research assistance.

2 For an informal discussion of the issue, see Norman Swan's interview with one of the researchers on 5 March 2007, available at http://www.abc.net.au/rn/healthreport/stories/2007/1861068.htm. The researchers were at pains to point out that their findings should not be extrapolated to foods that are rich in vitamins, such as fresh fruit and vegetables.

they drew together all the best evidence in the world, which in this case meant randomised trials of vitamins A, C and E, beta carotene and selenium, and found no evidence that vitamins make you live longer. If anything, those who took vitamin supplements seemed to live shorter lives.

Not wanting to send myself to an early grave, I stopped taking multivitamin pills.

What makes me so sure about this decision? First, because the evidence comes from a study published in one of the world's leading medical journals. Second, because medicine has a well-established hierarchy of evidence. Grade I evidence in that hierarchy is 'well-conducted systematic reviews of randomised trials'. (Grade II is non-randomised controlled trials and uncontrolled experiments, while Grade III is descriptive studies and case reports.) There is a strong consensus in the medical profession that when it comes to questions like 'Are vitamins good for you?', the highest grade of evidence is a systematic review of randomised trials.

Yet suppose I were a policy maker, charged with deciding whether to scrap or expand a social program. In many cases, I would probably find that the only available evidence was based upon anecdotes and case studies — what medical researchers would regard as the lowest grade of evidence in their hierarchy. While the evidence base in social policy is steadily advancing, a lack of data and an unwillingness to experiment are two major factors that hamper our understanding of what works and what does not.

Uncertainty over the effectiveness of our social policies is even more striking when you realise that annual government social expenditure amounts to around $8000 per Australian.[3] Ask a handful of experts, and you will find no shortage of ideas for improving the system. More education, more work experience, conditional cash transfers, laptop rollouts, higher income support, wage subsidies, lower minimum wages, public works programs and prison reform are among the recent favourites. Knowing more about which of these policies work, and why, could help us to improve social outcomes, save money or both.

10.2 The counterfactual problem

The challenge in assessing any policy intervention is that we need to know the counterfactual: what would have happened in the absence of the policy. If you are a farmer who is experimenting with a new fertiliser, this is pretty straightforward: put the fertiliser on every other plant, and the counterfactual is the unfertilised plants.

[3] This calculation uses the OECD definition of 'social expenditure', which amounted to $165 billion in 2005.

In the social sciences, this turns out to be a much tougher problem to address.[4] If we simply use time series variation, we may find it tricky to separate the policy change from secular changes in incidence over time. For example, say that we wanted to estimate the impact of a training program on earnings. If we were to just track the earnings of participants, we might find it difficult to separate the effect of the overall economy from the impact of the program.

Another approach is to construct a counterfactual by using those who choose not to participate as the control group. For example, suppose that you wanted to see the impact of alcohol management plans on outcomes in Indigenous communities. One evaluation strategy might compare outcomes in communities that chose to establish an alcohol management plan with outcomes in those that did not. But if the two sets of communities are systematically different — say, because the treatment group has more social capital or stronger leadership than the control group — then such an approach would be likely to overestimate the impact of the intervention.

10.3 The strengths of randomised trials

One way of getting around these problems is to conduct a randomised policy trial, in which participants are allocated to the treatment or control group by the toss of a coin. The beauty of randomisation is that, with a sufficiently large sample, the two groups are very likely to be identical, both on observable characteristics and on unobservable characteristics. Just as in a medical randomised trial of vitamin supplements, the only difference between the treatment and control groups is the intervention itself. So if we observe statistically significant differences between the two groups, we can be sure that they are due to the treatment and not to some other confounding factor.[5]

In Australian social policy, a canonical example of a randomised policy trial is the New South Wales Drug Court trial, conducted in 1999–2000. Offenders were referred to the Drug Court from local or district courts, underwent a detoxification program and were then dealt with by the Drug Court instead of a traditional judicial process. At the time it was established, the number of places in detoxification was limited, so participants in the evaluation were randomly assigned either to the treatment group (313 persons) or the control group (201 persons). They were then matched to court records in order to compare reoffending rates over the next year or

4 This is one of the reasons that I believe the social sciences should be known as the 'hard sciences' rather than the pejorative 'soft sciences'.

5 On randomised policy trials, see for example, The Economist (2002), Farrelly (2008) and Leigh (2003).

more. The evaluation found that the Drug Court was effective in reducing the rate of recidivism, and that while it was more expensive than the traditional judicial process, it more than paid for itself (Lind et al. 2002). At a recent conference celebrating the tenth anniversary of the Drug Court, speakers broadly acknowledged the role that the court has played in improving the lives of drug offenders and the general community (Knox 2009).

What is striking about the Drug Court trial is that it provides a ready answer to the shock jocks. Imagine the following exchange.

> Q: 'Minister, is it true that your program spends more on drug offenders? Why should taxpayers fork out more money to put drug addicts through detox programs, when we could spend less and just throw them in jail?'
>
> A: 'You bet we're spending more on this program, and that's because we have gold-standard evidence that it cuts crime. A year after release, those who went through the Drug Court were half as likely to have committed a drug offence, and less likely to have stolen. It's probably the case that Drug Courts help addicts kick the habit. But even if you don't care a whit about their wellbeing, you should be in favour of Drug Courts because they keep you and your family safe.'

In the case of the Drug Court, many of us probably had an expectation that the policy would reduce crime. But high-quality evaluations do not always produce the expected result. Staying for a moment with criminal justice interventions, take the example of 'Scared Straight', a program in which delinquent youth visit jails to be taught by prison staff and prisoners about life behind bars. The idea of the program — originally inspired by the 1978 Academy Award winning documentary of the same name — is to use exposure to prison to frighten young people away from a life of crime. In the 1980s and 1990s, several US states adopted Scared Straight programs.

Low-quality evaluations of Scared Straight, which simply compared participants with a non-random control group, had concluded in the past that such programs worked, reducing crime by up to 50 percent. Yet, after a while, some US states began carrying out rigorous randomised evaluations of Scared Straight. The startling finding was that Scared Straight actually increased crime, perhaps because youths discovered jail was actually not as bad as they had thought. It was not until policy makers moved from silver-standard evidence to gold-standard evidence that they learned the program was harming the very people it was intended to help (Boruch and Rui 2008; Petrosino et al. 2002).

Being surprised by policy findings is perfectly healthy. Indeed, we should be deeply suspicious of anyone who claims that they know what works based only on theory or small-scale observation. As economist John Maynard Keynes once put it in a

different context, ‘When the facts change, I change my mind. What do you do, sir?’[6]

10.4 Are randomised trials ethical?

Although it would be practically possible to randomly trial many of our social policy interventions, there is a reluctance among policy makers to subject policy interventions to gold-standard evaluation. In most developed nations, it is impossible to get a new pharmaceutical licensed without a randomised trial. Yet new social policies — often costing considerably more — require no such evaluation.

One common explanation proffered for this is the ethical challenge: when you have a program that you think is effective, how can you toss a coin to decide who receives it? The simplest answer to this is that the reason we are doing the trial is precisely because we do not know whether the program works. The simplest exposition of this is ‘Rossi’s Law’ (named after sociologist Peter Rossi), which states: ‘The expected value for any measured effect of a social program is zero.’ If you believe Rossi’s Law, it mostly does not matter whether a given individual is allocated to the treatment group or the control group. Indeed, for some programs — such as Scared Straight — participants in the control group ended up better off than those in the treatment group.

Adam Gamoran, a professor at the University of Wisconsin-Madison, takes the ethical argument a little further. If you know for sure whether a program works, Gamoran argues, then it is unethical to conduct a randomised trial. But if you do not know whether the program works, then it is unethical *not* to conduct a randomised trial. Every dollar we spend on an ineffective program is a dollar that could have been directed to a better program or returned to taxpayers. The quicker we can find out what works and what does not, the sooner we can direct resources to where they are needed most.

We should not lightly dismiss ethical concerns about randomised trials, but they are often overplayed. Medical researchers, having used randomised trials for several decades longer than social scientists, have now grown relatively comfortable with the ethics of randomised trials. Certain medical protocols could be adapted by social scientists, such as the principle that a trial should be stopped early if there is clear

6 Reply to a criticism during the Great Depression of having changed his position on monetary policy, as quoted in Malabre (1994, p. 220).

evidence of harm, or the common practice of testing new drugs against the best available alternative.

One example, again from New South Wales, helps to illustrate how much further advanced medical researchers are when it comes to randomised trials. For the past three years, an NRMA CareFlight team, led by Alan Garner, has been running the Head Injury Retrieval Trial (HIRT), which aims to answer two important questions: Are victims of serious head injuries more likely to recover if we can get a trauma physician onto the scene instead of a paramedic? And can society justify the extra expense of sending out a physician, or would the money be better spent in other parts of the health system?

To answer these questions, Garner's team is running a randomised trial. In effect, when a Sydney 000 operator receives a report of a serious head injury, a coin is tossed. Heads, you get an ambulance and a paramedic. Tails, you get a helicopter and a trauma physician. Once 500 head injury patients have gone through the study, the experiment will cease and the results will be analysed.

When writing an article about the trial last year, I spoke with Alan Garner, who told me that, although he has spent over a decade working on it, even he does not know what to expect from the results (Leigh 2008). 'We think this will work', he told me in a phone conversation, 'but so far, we've only got data from cohort studies.' Indeed, he even said, 'Like any medical intervention, there is even a possibility that sending a doctor will make things worse. I don't think that's the case, but [until HIRT ends] I don't have good evidence either way.'

For anyone who has heard policy makers confidently proclaim their favourite new idea, what is striking about Garner is his willingness to run a rigorous randomised trial, and listen to the evidence. Underlying HIRT is a passionate desire to help head injury patients, a firm commitment to the data and a modesty about the extent of our current knowledge.

10.5 The limits of randomised trials

While randomisation is an underused tool in the policy drawer, it is not effective in all cases. Writing tongue-in-cheek in the *British Medical Journal*, Gordon Smith and Jill Pell (2003) argued that the quality of the evidence on parachute effectiveness was severely limited by the absence of randomised controlled trials. They pointed out that the only evidence that parachutes prevent deaths when people jump out of planes was based on observation and expert opinion, the lowest rank of

evidence in the hierarchy. Their conclusion: we need randomised trials of parachutes!

As the previous section noted, randomised trials often need to undergo ethical scrutiny. Just as we would not allow a randomised trial of parachutes, we would not countenance a randomised trial that withdrew all income support from lone parents, experimented with taking children out of school or doubled hospital waiting lists. The *randomistas* — as economist Angus Deaton has called them — are not welcome everywhere.

But it is rare that policy makers are actually countenancing a policy contraction or expansion this radical. More often, the kinds of questions that need to be answered are much more modest, and therefore readily amenable to randomisation. Do long-term unemployed youth benefit more from job training or wage subsidies? Do schoolchildren benefit more from teacher merit pay or class size reductions? Would more generous post-release payments prevent ex-prisoners falling back into bad habits? What kinds of intensive early childhood programs would work best for Indigenous children?

Another limit to what we can learn from randomised trials comes from scale effects. As anyone who has eaten cafeteria food knows, what works well on a small scale does not necessarily work on a large scale (Currie 2001). One problem is that small-scale programs are often 'boutique programs', which are resourced to a level that is not feasible if implemented across an entire system. Another risk is that small-scale programs might fail to measure spillover and displacement effects. In economic jargon, randomised trials are a very precise way of measuring partial equilibrium effects, but often do not allow us to get at the general equilibrium effects.[7]

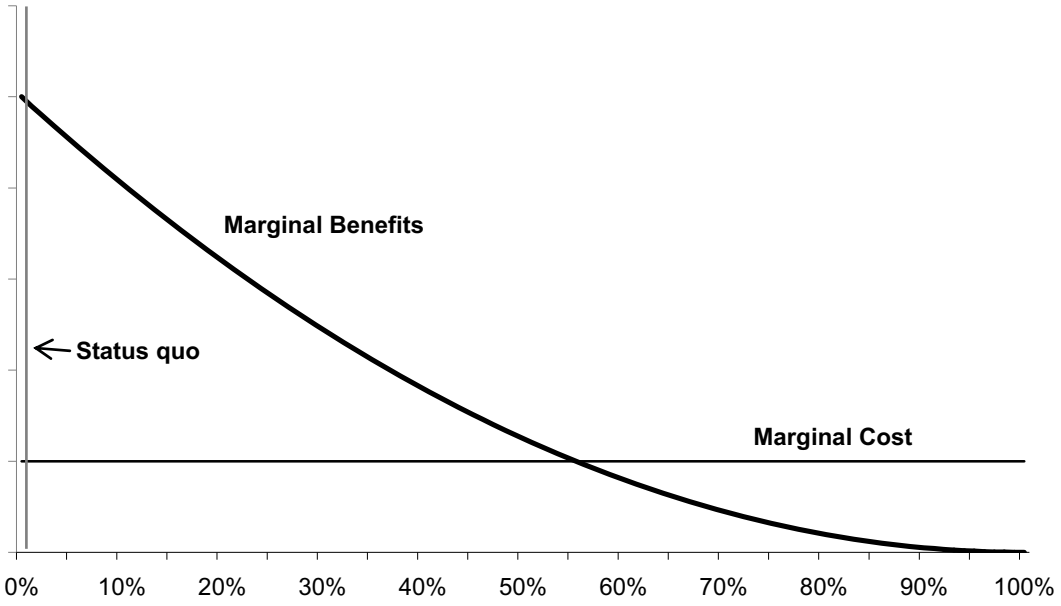
Because of these limitations, it is unlikely that we would ever want 100 per cent of government evaluations to be randomised trials. Most likely, the marginal benefit of each new randomised trial is a little lower than that of the previous one. At some point, it is indeed theoretically possible that we could end up doing more randomised trials than is socially optimal.

However, this is unlikely to ever occur, at least in my lifetime. Figure 10.1 presents my stylised sense of the state of play for randomised trials in Australia. My best

7 For a thoughtful discussion of these issues in the context of international development, see the papers presented at a Brookings Global Economy and Development Conference 'What Works in Development? Thinking Big and Thinking Small', held in Washington DC on 29–30 May 2008. Papers and presentations are available at http://www.brookings.edu/events/2008/0529_global_development.aspx. See also the recent exchange in Deaton (2009), Heckman and Urzua (2009) and Imbens (2009).

estimate is that less than 1 per cent of all government evaluations are randomised trials (excluding health and traffic evaluations, the proportion is probably less than 0.1 per cent).[8] Another way to put this is that, *to a first approximation, Australia currently does no randomised policy trials*. Policy makers could safely embark on a massive expansion of randomised policy trials in Australia before we come close to the point where the costs exceed the benefits.

Figure 10.1 **Costs and benefits of more randomised policy trials**



Note: All datapoints in this chart are assumptions.

8 I was unable to find a comprehensive database of all government evaluations in Australia, so opted instead to conduct a Google search on the gov.au domain. A search on 14 August 2009 for 'randomised AND evaluation' brought up fewer than 10,000 hits, while a search for just 'evaluation' brought up more than 1.7 million hits. It is likely that both these numbers overestimate the true numbers of randomised and non-randomised evaluations, since they are counts of hits rather than unique files. However, to the extent that the ratio of hits to unique files is the same for randomised and non-randomised evaluations (which seems a reasonable assumption), this suggests that about 0.5 percent of Australian government evaluations in recent years have used a randomised design. The assumption about most Australian randomised trials being health and traffic evaluations is based on the fact that the Australian trials contained within the Campbell Collaboration's randomised trials register (C2-SPECTR) are largely in those two categories (see Leigh 2003 for cross-national comparisons of the contents of this register).

There are also limits to the power of good evidence to change the minds of policy makers. In a recent speech, Gary Banks catalogued several instances — ranging from fertility policy to industry policy — in which the Productivity Commission showed that policies had failed to achieve their goal (Banks 2009). Despite the rigour of the econometrics, some readers may be surprised to learn that the Commission's recommendations have not been universally adopted by the Australian Government. Nonetheless, raising the evidence bar matters. High-quality evaluations are harder for policy makers to dismiss than low-quality evaluations.

10.6 Promoting randomised trials

How might randomised trials be promoted in Australia? One possibility would be for the federal government to systematically set aside resources for states to conduct rigorous randomised trials that would have a national benefit. For example, despite spending billions of dollars to reduce class sizes over the past few decades, Australia has never conducted a randomised trial of the impact of smaller classes on student performance. Part of the reason for this is political: state politicians are uncomfortable telling voters that a lottery will be used to determine which students get to sit in the smaller class. But if the program were federally funded, it is possible that a state government might be willing to administer the experiment. In cases where federal programs are being channelled through the states, small amounts set aside for 'random assignment evaluation' can have a large knowledge payoff. Box 10.1 sets out three examples from recent pieces of US federal legislation that explicitly set aside funds for randomised evaluation.[9]

[9] These examples are drawn from the Coalition for Evidence-Based Policy, which is part of the Council for Excellence in Government (http://www.excelgov.org/), and a presentation by Adam Gamoran, 'Measuring impact in science education: challenges and possibilities of experimental design', NYU Abu Dhabi Conference, January 2009.

Box 10.1 US federal legislation that specifically funds randomised evaluation

1. The Second Chance Act, dealing with strategies to facilitate prisoner re-entry into the community, sets aside 2% of program funds for evaluations that 'include, to the maximum extent possible, random assignment … and generate evidence on which re-entry approaches and strategies are most effective'.
2. The No Child Left Behind Act calls for evaluation 'using rigorous methodological designs and techniques, including control groups and random assignment, to the extent feasible, to produce reliable evidence of effectiveness.'
3. Legislation to improve child development via home visits directs the Department of Health and Human Services to 'ensure that States use the funds to support models that have been shown in well-designed randomized controlled trials, to produce sizeable, sustained effects on important child outcomes such as abuse and neglect'.

Another way that randomised trials might be promoted is through the regular use of an 'evidence hierarchy' by social policy makers. Such hierarchies — common in the medical literature — are steadily gaining currency among policy makers. As an example, Box 10.2 depicts the hierarchy proposed in Leigh (2009). Although evidence hierarchies are invariably imperfect, they can help to focus attention on the quality of the available knowledge base. If a politician is told 'we think you should implement Option A, but you should also know that the state of knowledge is very poor', he or she may be more inclined to sow the seeds of a few new randomised trials.

Box 10.2 A possible evidence hierarchy for Australian policy makers

1. Systematic reviews (meta-analyses) of multiple randomised trials
2. High-quality randomised trials
3. Systematic reviews (meta-analyses) of natural experiments and before–after studies
4. Natural experiments (quasi-experiments) using techniques such as differences-in-differences, regression discontinuity, matching or multiple regression
5. Before–after (pre–post) studies
6. Expert opinion and theoretical conjecture

All else equal, studies should also be preferred if they are published in high-quality journals, if they use Australian data, if they are published more recently and if they are more similar to the policy under consideration.

Source: Leigh (2009).

10.7 Conclusion

In Australian policy debates, the term 'evidence-based policy making' has now become so meaningless that it should probably be jettisoned altogether. The problem in many domains is not that decision makers do not read the available literature — it is that they do not set up policies in such a way that we can learn clear lessons from them.[10] In employment policy, the early 1990s recession saw Australia spend vast amounts on active labour market programs, without producing a skerrick of gold-standard evidence on what works and what does not. In Indigenous policy, there are as many theories as advocates but precious few randomised experiments that provide hard evidence about what really works.

Sometimes randomised trials will justify the expansion of a politically difficult intervention. But we should never forget the social benefit of evaluations whose results show us that a program does not work. Thanks to randomised evaluations of multivitamin tablets, my household now has $30 a year to spend on other things. The same is true of evaluations that find government programs to be ineffective. A randomised trial that conclusively shows a program had no impact is a valuable piece of knowledge in improving Australian public policy.

References

Banks, G. 2009, 'Evidence-based policy-making: What is it? How do we get it?', lecture given at the *ANZSOG/ANU Public Lecture Series*, Canberra, 4 February 2009, http://www.pc.gov.au/__data/assets/pdf_file/0003/85836/cs20090204.pdf (accessed 3 August 2009).

Bjelakovic, G., Nikolova, D., Gluud, L.L., Simonetti, R.G. and Gluud, C. 2007, 'Mortality in randomized trials of antioxidant supplements for primary and secondary prevention: systematic review and meta-analysis', *Journal of the American Medical Association*, vol. 297, no. 8, pp. 842–857.

Boruch, R. and Rui, N. 2008, 'From randomized controlled trials to evidence grading schemes: current state of evidence-based practice in social sciences', *Journal of Evidence-Based Medicine*, vol. 1, no. 1, pp. 41–49.

Currie, J. 2001, 'Early childhood education programs', *Journal of Economic Perspectives*, vol. 15, no. 2, pp. 213–238.

10 Harvard economist Roland Fryer is particularly scathing about the quality of evidence in education. In a recent interview with the *New York Times*, he said: 'If the doctor said to you, "You have a cold; here are three pills my buddy in Charlotte uses and he says they work," you would run out and find another doctor. Somehow, in education, that approach is O.K.' (quoted in Hernandez 2008).

Deaton, A.S. 2009, 'Instruments of development: randomization in the tropics, and the search for the elusive keys to economic development', NBER Working Paper no. 14690, http://www.nber.org/papers/w14690 (accessed 3 August 2009).

Farrelly, R. 2008, 'Policy on trial', *Policy*, vol. 24, no. 3, pp. 7–12.

Heckman, J. and Urzua, S. 2009, 'Comparing IV with structural models: what simple IV can and cannot identify', NBER Working Paper no. 14706, http://www.nber.org/papers/w14706 (accessed 3 August 2009).

Hernandez, J.C. 2008, 'New effort aims to test theories of education', *New York Times*, 25 September.

Imbens, G.W. 2009, 'Better LATE than nothing: some comments on Deaton (2009) and Heckman and Urzua (2009)', NBER Working Paper no. 14896, http://www.nber.org/papers/w14896 (accessed 3 August 2009)

Knox, M. 2009, 'Applause for former drug users who turn their lives around', *Sydney Morning Herald*, 7 February.

Leigh, A. 2003, 'Randomised policy trials', *Agenda: A Journal of Policy Analysis and Reform*, vol. 10, no. 4, pp. 341–354.

—— 2008, 'A good test of public policy', *Australian Financial Review*, 8 April.

—— 2009, 'What evidence should social policymakers use?', *Australian Treasury Economic Roundup*, vol. 1, pp. 27–43.

Lind, B., Weatherburn, D., Chen, S., Shanahan, M., Lancsar, E., Haas, M. and De Abreu Lourenco, R. 2002, *NSW Drug Court evaluation: cost-effectiveness*, NSW Bureau of Crime Statistics and Research, Sydney, www.courtwise.nsw.gov.au /lawlink/bocsar/ll_bocsar.nsf/vwFiles/L15.pdf/$file/L15.pdf (accessed 3 August 2009).

Malabre, A.L. 1994, *Lost Prophets: An Insider's History of the Modern Economists*, Harvard Business Press, Cambridge, Massachusetts.

Petrosino, A., Turpin-Petrosino, C. and Buehler, J. 2002, *'Scared Straight' and other juvenile awareness programs for preventing juvenile delinquency* (Updated C2 Review), Campbell Collaboration Reviews of Intervention and Policy Evaluations (C2-RIPE), available at www.campbellcollaboration.org.

Smith, G.C.S. and Pell, J.P. 2003, 'Parachute use to prevent death and major trauma related to gravitational challenge: systematic review of randomised controlled trials', *British Medical Journal*, vol. 327, pp. 1459–1461.

The Economist 2002, 'Try it and see', 2 March, pp.73–74.