Australian Economic Review

SPECIAL ARTICLE

The Case for Randomised Trials (and Why Big Data Does Not Supersede Randomisation)

Andrew Leigh

Parliament of Australia, Canberra, Australia

Correspondence: Andrew Leigh (andrew.leigh.mp@aph.gov.au)

Received: 13 October 2024 | **Revised:** 19 February 2025 | **Accepted:** 25 February 2025

Keywords: evaluation | field experiments | global evidence sharing | institutional barriers to evaluation | randomised controlled trials | randomised trial infrastructure | selection bias in observational data

ABSTRACT

Research Question/Issue: With the growing availability of large-scale datasets, is randomisation still necessary for identifying causal impacts?

Research Findings/Insights: Randomised trials, by using luck to assign participants to treatment and control groups, reliably provide a credible counterfactual that ensures observed differences reflect causal impacts. In contrast, observational data often produces misleading correlations that fail to replicate under experimental conditions. Therefore, the increased availability of big data does not make randomisation obsolete.

Practitioner/Policy Implications: I propose five approaches to increase the quality and quantity of randomised policy trials: encourage curiosity in yourself and those you lead; seek simple trials, especially at the outset; ensure experiments are ethically grounded; foster institutions that push people towards more rigorous evaluation; and collaborate internationally to share best practice and identify evidence gaps.

Methods Used: This paper employs a qualitative synthesis of historical and contemporary examples, illustrating the superiority of randomised trials over purely observational methods. By drawing comparisons across disciplines—economics, health, and social policy—it highlights how nonexperimental approaches can fall short and explores how big data can be a complement to rigorous randomised trials.

JEL Classification: D04, H43, C93

1 | Introduction

In 1747, 31-year-old Scottish naval surgeon James Lind set about determining the most effective treatment for scurvy, a disease that was killing thousands of sailors around the world. Selecting 12 sailors suffering from scurvy, Lind divided them into six pairs. Each pair received a different treatment: cider; sulphuric acid; vinegar; seawater; a concoction of nutmeg, garlic and mustard; and two oranges and a lemon. In less than a week, the pair who had received oranges and lemons were back on active duty, while the others languished. Given that sulphuric acid was then the British Navy's main treatment for scurvy, this was a crucial finding.

The trial provided robust evidence for the powers of citrus because it created a credible counterfactual. The sailors did not choose their treatments, nor were they assigned based on the severity of their ailment. Instead, they were randomly allocated, making it likely that differences in their recovery were due to the treatment rather than other characteristics.

This article expands on a speech delivered at the University of Oxford on 2 October 2024, which was organised by Philip Clarke and hosted by Oxford Population Health's REAL Supply and Demand Units, the Oxford Health Economics Research Centre, and the Oxford Centre for Health Economics.

Lind's randomised trial, one of the first in history, has attained legendary status. Yet because 1747 was so long ago, it is easy to imagine that the methods he used are no longer applicable. After all, Lind's research was conducted at a time before electricity, cars and trains, an era when slavery was rampant and education was reserved for the elite. Surely, some argue, ideas from such an age have been superseded today.

In this article, I make the case for randomised policy trials and discuss how the architecture supporting randomised trials might be strengthened. Recent decades have seen a substantial increase in the use of randomised evaluations to assess program impact, including across welfare, employment, education and crime (Baron 2018). Over the first two decades of the 21st century, the share of papers in the top five economics journals that were on field experiments quadrupled (List 2024). The awarding of the 2019 Economics Nobel Prize to Abhijit Banerjee, Esther Duflo and Michael Kremer reflected the prominence of randomised trials. At the same time, the infrastructure supporting randomised trials has grown through government initiatives such as the UK What Works Centres and philanthropic initiatives such as the US Coalition for Evidence-Based Policy. Both can now point to a body of social programs that have been shown by large-scale randomised trials to produce a meaningful improvement in economic opportunity.

The remainder of the article is structured as follows. Section 2 discusses why large data sets do not preclude randomised trials, drawing on examples from health and the social sciences. Section 3 explores the growth of randomised trials and the extent to which medical randomised trials have outpaced other fields. Turning from challenges to solutions, Section 4 outlines five approaches to fostering randomised trials. The final section concludes.

2 | Big Data Does Not Make Randomised Trials Obsolete

In place of randomised trials, some put their faith in 'big data'. Between large-scale surveys and extensive administrative data sets, the world is awash in data as never before. Each day, hundreds of exabytes of data are produced. Big data has improved the accuracy of weather forecasts, permitted researchers to study social interactions across racial and ethnic lines, enabled the analysis of income mobility at a fine geographic scale and much more.

Yet a clue to the value of randomised trials comes from the behaviour of the biggest big data company of them all, Google. Since its founding in 1998, Google has conducted thousands of randomised trials to refine its products. The company regularly conducts randomised trials (often dubbed A/B testing) to see how users prefer search results to be spaced out, which colours to use, and whether new features should be added to the product (see e.g. Gomes 2008; Christian 2012).

Why would Google conduct randomised trials rather than using big data? Because it is keen to uncover causal effects. To see this, suppose that the company instead decided to determine the impact of product tweaks by looking at patterns in the data. For example, it could offer a new function in Google Sheets, and compare the productivity of users who took it up with the productivity of users who did not take it up. Such an analysis might also hold constant other observed factors about the two groups of users, such as how often they use the product.

The problem with such an analysis is that what is not observed can have a major impact on productivity. If users who like new functions are increasing their productivity at a more rapid rate, then this will bias the estimate upwards. Conversely, if users who like new functions are procrastinating, it will bias the estimate downwards. Google does not know the true answer, so it opts for a randomised trial. In conducting its randomised trials, big data is a massive asset for Google. But big data does not preclude the need to do randomised trials.

Another example arises in heart health (Collins et al. 2020). Randomised trials have demonstrated a strongly beneficial effect of statins on reducing cardiovascular mortality. Yet when they analysed a database covering the entire Danish population, researchers found that the chance of death from cardiovascular causes was one-quarter higher among those who took statins than among those who did not. The explanation is straightforward: people who were prescribed statins were at elevated risk of having a heart attack. Yet even when researchers made statistical adjustments, using all the variables available in the database, they were unable to reproduce the well-known finding that statins have a beneficial effect on cardiovascular mortality.

Analysis of the Danish database also suggested that the relative risk of cancer was 15% lower among patients who took statins, an effect that remained statistically significant even after controlling for other observed factors about the patients. Yet this result is at odds with the evidence from randomised trials. A meta-analysis of randomised trials, covering more than 10,000 cases of cancer, found no effects of statins on the incidence of cancer, nor on deaths from cancer. On average, these randomised trials covered a 5-year period; longer than in the non-randomised database analysis.

The observational data was doubly wrong. Observational data failed to replicate the well-known finding that statins improve heart health. And observational data wrongly suggested that statins reduce the risk of cancer. Randomised trials, which were not biased by selection effects, provided the correct answer.

A similar issue arose with estimating the health impact of hormone replacement therapy for postmenopausal women. In 1976, the Nurses' Health Study began tracking over 100,000 registered nurses. The study found that women who chose to use hormone replacement therapy halved their risk of heart disease (Stampfer et al. 1985). By the late 1990s, around two-fifths of postmenopausal women in the United States were using hormone replacement therapy—mostly to reduce the risk of heart disease. However, no randomised trial had evaluated the impact of hormone replacement therapy.

Then the National Institutes of Health funded two randomised trials, comparing hormone replacement therapy against a placebo (Manson et al. 2024). The trials, which began in 1993, did not support menopausal hormone replacement therapy to

prevent coronary heart disease. Indeed, one of the randomised trials was stopped early because the data and safety monitoring board concluded that there was some evidence of harm. With the health of millions of women at stake, the early observational data had presented an inaccurate picture of the impact of hormone replacement therapy. The fact that the observational studies had a larger sample size than the randomised trials did not help. Lacking evidence from randomised trials, millions of women took a treatment that had an adverse impact on their health. Randomised trials uncovered the truth.

Researcher Rory Collins and his co-authors refer to this as the 'magic of randomisation' (Collins et al. 2020). Large data sets are a valuable *complement* to randomised trials. But big data is not a *substitute* for randomisation.

In a 2023 joint statement, the European Society of Cardiology, American Heart Association, American College of Cardiology, and the World Heart Federation (Bowman et al. 2023) concluded that 'The widespread availability of large-scale, population-wide, real-world data is increasingly being promoted as a way of bypassing the challenges of conducting randomized trials. Yet, despite the small random errors around the estimates of the effects of an intervention that can be yielded by analyses of such large data sets, non-randomized observational analyses of the effects of an intervention should not be relied on as a substitute, due to their potential for systematic error'.

In their statement, the four cardiology organisations call for measures to ensure that randomised trials are 'fit for the twenty-first century', addressing issues such as rising cost and complexity. In addition, well-conducted randomised trials should have a sufficient sample size to detect a meaningful effect if one exists; should have low attrition; should be conducted in a manner that is representative of the scaled-up program; and should be pre-registered.

While correlations in large data sets do not necessarily indicate causation, administrative data can be enormously helpful in ensuring the precision of estimates from randomised trials. The advantage of administrative data in this context was illustrated by the Oregon Health Insurance Experiment, which studied the impact of Medicaid coverage for uninsured low-income adults on emergency room use. While survey data showed no statistically significant effect, administrative data revealed a 40% increase (Taubman et al. 2014). This discrepancy was partly due to the greater accuracy of administrative records. Even when restricted to the same individuals and time periods, administrative data produced larger, more precise estimates. Additionally, it covered a longer period and provided richer details on visit timing, diagnoses, and usage history— information difficult to obtain through surveys.

Finkelstein and Taubman (2015b) note further advantages of administrative data for randomised trials: they are easier and cheaper to obtain; they include a near-census of relevant individuals; they are less likely than survey data to suffer from the possibility of bias in favour of what the researchers would like to hear; and they can be useful for following up on long-term outcomes. For an overview of the main government administrative data sources for evaluation in Australia, see Australian Centre for Evaluation (2025).

3 | The Growth of Randomised Trials

In policymaking, randomised trials have been deployed in unexpected places. Randomised trials of policing strategies have shown that hot spots policing reduces crime (Sherman and Weisburd 1995). A randomised trial of incarceration policies found that releasing prisoners 6 months early did not raise recidivism (Berecochea and Jaman 1981). A randomised trial found that when people in India were given a financial incentive to get their licence earlier, they were more likely to bribe the tester (Bertrand et al. 2007). A randomised trial in Mexico found that road upgrades boosted property prices and reduced poverty (Gonzalez-Navarro and Quintana-Domeque 2016). A randomised trial with airline pilots found that providing feedback on fuel use led captains to be more economical, saving the airline a million litres of fuel (Gosnell et al. 2016). Economists are also integrating the findings from randomised trials into macroeconomic models (Buera et al. 2023), using field experiments to carefully test economic theories (Banerjee 2020), and combining evidence from field and natural experiments for parameter estimation (Bergquist et al. 2022).

Yet by comparison with health, the uptake of randomised trials in the social sciences remains modest. In the 1970s, there were fewer than 10 randomised trials published each year in either health or the social sciences. By the 2020s, around 24,000 health randomised trials were published each year, compared with around 1300 social sciences randomised trials (Campbell Collaboration analysis of Web of Science data, published in Halpern and Maru 2024). For every published randomised trial in the social sciences, there are over fifteen in health.[1] This seems lop-sided given the breadth of the social sciences, covering welfare, education, crime, housing, aged care and more. Combined spending across these social science policy areas exceeds spending on health. Yet in terms of randomised trials, health is in a league of its own.

Past Australian Governments have devoted significant resources towards social initiatives that are intended to test new approaches, but which lack a suitable evaluation framework. Part of the challenge arises from the fact that programs have often been tailored to the intervention group, making it difficult to draw general conclusions about what works and what does not. Where communities volunteer to develop an intervention, it can be near-impossible to separate the impact of the program from other aspects of the community itself.

Two large-scale funding programs illustrate this challenge. From 2000 to 2009, the Australian Government spent over $500 million on the Stronger Families and Communities Strategy. The strategy included 'Local Solutions to Local Problems', which provided funding for up to 500 communities for locally developed solutions, and the 'Stronger Families Fund', which funded up to 75 projects, developed by First Nations communities. The bespoke nature of the interventions and the lack of a credible control group meant that two major evaluations of the strategy were unable to draw strong conclusions about what worked and what did not.

From 2016 to 2020, the Australian Government spent $96 million on a Try, Test and Learn Fund, which supported 52 projects aimed at assisting groups thought to be at risk of long-term welfare dependency. Like the evaluations of the Stronger Families and Communities Strategy, the evaluation of the Try, Test and Learn Fund was quasi-experimental, based on comparing outcomes in treated groups with similar groups who did not receive the intervention. As the evaluation report noted, 'The gold standard, which is commonly used in medical trials, is to randomise project allocation... In Try, Test and Learn and other projects where participation is based on self-selection, impact analysis involves the construction of a comparison group from nonparticipants' (Institute for Social Science Research 2021, 118). Although the evaluation of the fund used administrative data, it acknowledged that participants who were recruited to the program might have 'differed from the comparison group in ways that could not be observed but could impact on client outcomes' (Institute for Social Science Research 2021, 8). Despite its size, the fund has not significantly contributed to identifying the most effective programs for breaking the cycle of disadvantage.

The Thodey Review of the Australian Public Service concluded that in-house evaluation capacity had declined, citing research that a lack of evaluation expertise 'diminishes accountability and is a significant barrier to evidence-based policy-making' (Thodey et al. 2019, 221). A study from the think tank CEDA examined a sample of 20 Australian Government programs conducted between 2015 and 2022 (Winzar et al. 2023). The programs had a total expenditure of over $200 billion. CEDA found that 95% were not properly evaluated. CEDA's analysis of state and territory government evaluations reported similar results. As the CEDA researchers note, 'The problems with evaluation start from the outset of program and policy design'. Across the board, CEDA estimates that fewer than 1.5% of Australian government evaluations use a randomised design (Winzar et al. 2023, 44).

The relatively small number of randomised trials of social programs is particularly troubling given what the evidence tells us about the programs that are rigorously evaluated. In health, only one in ten drugs make it through Phase I, II and III clinical trials and onto the market (Hay et al. 2014). In education, an analysis of randomised trials commissioned by the US Department of Education's Institute of Education Sciences found that only one in ten generated the intended effects (Coalition for Evidence-Based Policy 2013). And remember all those trials that Google is doing? Google estimates that just one in five randomised trials help refine the company's applications (Thomke 2013).

4 | Five Strategies for Strengthening Randomised Trials

This suggests that the best approach to policymaking is what US President Franklin D. Roosevelt once called 'bold, persistent experimentation' (Roosevelt 1932). If many promising policies do not work as well as intended, then rigorous evaluation is essential to building a cycle of continuous improvement. Rigorous evaluation guarantees that government policies in a decade's time will be more effective than they are today. A failure to evaluate runs the risk that we will unwittingly repeat our mistakes. Evaluation puts us in a virtuous feedback loop. Without it, we can end up in a doom loop.

How can we encourage more rigorous evaluation? There are five approaches that can promote more high-quality evaluations, especially randomised trials.

First, encourage curiosity. Employees quickly come to understand the culture of an organisation. Some organisations foster questioning; others favour conformity. Encouraging a culture of curiosity does not come naturally to many managers. Questions can be perceived as time-consuming or distracting. Yet when managers make clear that they value new insights, they give permission for everyone in the organisation to question accepted wisdom and gather better evidence.

David Cowan, who has run multiple randomised trials of policing programs in Victoria, notes that managers can enable randomised trials by simply asking questions such as 'do we know if it works?' and 'how could we find out?' (Mazerolle et al. 2021). Strong organisations encourage the philosophy of what the UK Behavioural Insights Team famously dubbed 'Test-Learn-Adapt' (Haynes et al. 2012). This accords with the advice offered by Briscese and List (2024), who suggest that one way to boost the uptake of randomised policy trials is to educate the public on the value of open-mindedness.

Second, favour simplicity. Some of the most famous randomised trials are among the most complex. The Negative Income Tax experiments of the 1970s, the Job Training Partnership Act experiments of the 1980s and the Moving to Opportunity experiment of the 1990s each cost millions of dollars. These experiments have had a major impact on public policy, but a side-effect is that they have left some people with the mistaken impression that randomised trials must always be costly and time-consuming. Yet many experiments can be quick and simple. In New South Wales, a randomised experiment showed that when a red 'Pay Now' stamp was added to traffic fines, repayment rates rose from 14% to 17%, delivering $10 million a year in state revenue (Halpern 2015, 89). Another randomised experiment sent letters to Australian general practitioners who appeared to be excessively prescribing antibiotics, pointing that they were prescribing at higher rates than most other doctors. Doctors who received the letter wrote 12% fewer prescriptions—significantly helping to address the problem of antimicrobial resistance (Behavioural Economics Team of the Australian Government 2018).

Government officials charged with sending out letters, emails or text messages should have the functionality to send two versions, so they can continuously improve the language and messaging of their correspondence. This kind of A/B testing has been standard for market research companies for decades, yet remains rare in the public sector. Simple randomised trials—evaluating a tweak rather than a new initiative— are also more likely to change practice. An analysis of 73 randomised trials conducted by cities in conjunction with the BIT-North America Nudge Unit found that trials which took place in the context of ongoing communication to citizens (such as altering a mailer about business tax registration) were highly likely to lead to a

change in practice after the trial ended (DellaVigna et al. 2024). Such randomised trials had an adoption rate of 67%, significantly higher than the 12% adoption rate for randomised trials of new interventions.

Another initiative is grant funding to support low-cost randomised trials. In 2024, the Paul Ramsay Foundation, Australia's largest charitable foundation, issued a $2.1 million call for proposals for seven projects of up to $300,000 to be randomly evaluated. A similar approach has been adopted for the past decade by the Laura and John Arnold Foundation, who have funded a swath of low-cost trials across social policy.[2]

Third, ensure that all experiments are ethical. Subjecting randomised trials to appropriate ethical scrutiny is not just the right thing to do; it is also important for creating an environment in which further trials can be conducted. Ethical review ensures that the interests of vulnerable people are taken into account, and that the trial can be expected to improve overall wellbeing.

The ethical review process can also encourage researchers to think creatively about how to generate random variation in instances where researchers have strong prior beliefs that it will be effective. For example, suppose a program thought to have a positive effect was being rolled out nationwide over a 2-year period. A randomised trial might be conducted by randomising the regions that receive the program first. This is the approach taken by Muralidharan et al. (Forthcoming), who worked with the government of the Indian state of Jharkhand to randomise the order in which it introduced biometrically linked identity numbers across 132 sub-districts, covering 15 million people. This so-called 'stepped wedge' design produces rigorous causal impacts, while ensuring that everyone ends up with access to the program. Such a design allows a randomised trial to be conducted in instances where there is a political desire to deliver a universal program.

Another approach is to evaluate the impact of a universal program by advertising it to a randomly selected group of eligible people, thereby inducing differences in take-up. Such an approach, known as an 'encouragement design' was used by Finkelstein and Notowidigdo (2019), who promoted the Supplemental Nutrition Assistance Program to a group of people who were eligible but not enrolled. By inducing variation in take-up, they were able to show that the program had a positive causal impact on recipients' health and income. From an ethical perspective, it is worth noting that the control group were not denied access to the program; they simply failed to receive the promotional materials.

Fourth, build institutions to support evidence-based decision-making. Having bodies that promote high-quality evaluation can help to provide toolkits, seminars and nudges to inform and entice decision-makers towards rigorous evaluation approaches. Within a few years of its creation, the tiny UK Behavioural Insights Team ('the Nudge Unit') had carried out more randomised trials than the centre of the British Government had done in its entire history (Halpern 2015, 274). To date, three-fifths of schools in England have taken part in a randomised trial funded by the Education Endowment Foundation (Leigh 2025). The UK's What Works Centres (Sanders and Breckon 2023), Evaluation Taskforce and Magenta Book (HM Treasury 2020) have also provided a powerful impetus towards improving the quality and quantity of rigorous evaluations. In 2024, the Wellcome Trust and UK Economic and Social Research Council committed a combined £56 million towards the development of living evidence syntheses. This coincided with an announcement by three major organisations that work on evidence synthesis—JBI, Cochrane and Campbell—that they will work together to build a truly global evidence ecosystem.

In 2023, the Australian Government established the Australian Centre for Evaluation. Located within Treasury, the centre has a budget of around $2 million per year, and a staff of just over a dozen people. Its mandate is to 'put evaluation evidence at the heart of policy design and decision-making'. The main goal of the centre is to work collaboratively with government departments to conduct rigorous evaluations, especially randomised trials. Several trials are underway, including experiments to improve the quality of employment services. Implementing a recommendation of the Thodey Review to improve evaluation capacity across government, the Australian Centre for Evaluation has trained hundreds of public servants in evaluation and is presently preparing its first report on the state of evaluation across the public service.

Beyond centres, other institutional features that can help embed best practice evaluation include rules around the evidence that accompanies cabinet submissions and expectations about evaluation for programs with sunset clauses. For example, CEDA proposes that the Australian Government systematically review all programs with spending of more than $100 million at least every 5 years (Winzar et al. 2023, 42).

Another way that evaluation can be made more routine is if it becomes part of the expected toolkit for public sector managers. Just imagine how much the volume of randomised policy trials might grow if public servants know that they will eventually be asked in a promotion interview not only 'tell me about a workplace conflict you have resolved?' but also 'tell me about a randomised trial you have conducted?'.

Fifth, strengthen international collaboration. When researching my book *Randomistas*, Australian medical researcher David Johnson told me that in his field of kidney disease, he wants to see an end to underpowered single-centre trials. 'It's seldom that clinical practice will ever be changed by a single-centre trial', he argues. 'The numbers tend to be small, and even if they're big enough, you worry that they won't generalise' (Leigh 2018, 197). For his kind of medical research, Johnson contends, the future lies in coordinated research across multiple centres. Multi-country trials provided an inbuilt replication function, and greater assurance that interventions worked across people of different ethnicities. Multi-country trials can also reduce duplication of effort, making healthcare more efficient. Empirically, the trend towards international collaboration is borne out in the published research. From 1997 to 2019, the share of randomised trials in the medical literature that involved a multi-country team of researchers more than doubled, growing from 20% to 47% (Fukuhara et al. 2024).

Perhaps this approach has something to teach those of us seeking to better understand what works in policy. In their *Global Evidence Report: A Blueprint for Better International Collaboration on Evidence*, David Halpern and Deelan Maru advocate for countries to collaborate on evidence Living Evidence Reviews— research syntheses on key topics such as homelessness, job training or policing (Halpern and Maru 2024). They argue that such a collaboration could begin with the United Kingdom, United States, Canada and Australia. It would enable sharing of what works and what does not, as well as a recognition of the evidence gaps. The report also discusses other collaboration opportunities, such as a shared evaluation fund and international public service professional networks.

Other international institutions can play a role. In February 2025, Australia hosted an OECD workshop in Paris on 'Rigorous Impact Evaluation Approaches including Randomised Controlled Trials', at which speakers argued for the OECD to play a stronger role in improving the quality of evidence. The *Global Commission on Evidence* report (for which I served as a Commissioner) suggested that the World Bank devote a future *World Development Report* to how best to produce, share and use evidence (Global Commission on Evidence to Address Societal Challenges 2022).

5 | Conclusion

Large data sets can provide a wide range of insights into correlations across the population and can enable researchers to implement quasi-experimental approaches, such as regression discontinuity analysis. Yet the growing availability of data, even at a population-wide level, does not preclude randomised trials. Large administrative data sets are a boon to researchers conducting randomised trials, but merely analysing correlations in these data sets can mislead. Relying on big data without randomisation might lead to the mistaken conclusion that if millions of people in a city carry umbrellas, those umbrellas cause the downpour.

In policy, randomised trials have been used in a wide range of areas. In some cases, the results have suggested that interventions have extremely high payoffs. A German randomised trial found that a brochure informing at-risk job seekers about job search strategies increased earnings, with a benefit-cost ratio of 450:1 (Altmann et al. 2018). A Chicago randomised trial indicated that cognitive behavioural therapy for disadvantaged young men had a benefit-cost ratio of 30:1 (Heller et al. 2013). A randomised trial that helped low-income US parents complete university financial aid applications cost just US$650 per additional enrolled student (Bettinger et al. 2012). A randomised trial of messaging to taxpayers in the Dominican Republic boosted revenue by over US$100 million at virtually no cost (Holz et al. 2020).

Yet the volume of randomised trials in health dwarfs the volume of randomised trials in the social sciences. To increase the number of public policies subjected to randomised trials, I advocate five approaches. Encourage curiosity in yourself and those you lead. Seek simple trials, especially at the outset. Ensure experiments are ethically grounded. Foster institutions that push people towards more rigorous evaluation. Collaborate internationally to share best practice and identify evidence gaps.

It is worth noting that most of these suggested approaches apply to evaluation more broadly, not just randomised trials. As the Thodey Review noted, 'One challenge to reversing the decline in evaluation work across the Australian public service is the risk of failures... being exposed' (Thodey et al. 2019, 222). A more honest and transparent culture not only promotes democratic accountability, it also enables the evidence-based improvement of public services. A culture of evaluation, grounded in curiosity, integrity and international cooperation, allows government programs to become more efficient and effective. Without rigorous evaluation, policies risk being like astrology: comforting to some, but unlikely to withstand serious scrutiny.

James Lind's scurvy trial was pathbreaking. Alas, his writeup left something to be desired. Six years after his experiment, Lind published the 456-page tome *A Treatise of the Scurvy*. His experimental results were spot-on, but Lind's theoretical explanations for why citrus worked were hocus-pocus. The treatise was largely ignored.

Then, in the 1790s, a disciple of Lind, surgeon Gilbert Blane, was able to persuade senior naval officials that oranges and lemons could prevent scurvy. In 1795—almost half a century after Lind's findings—lemon juice was issued on demand. By 1799 it became part of the standard provisions. By the early 1800s, British naval sailors were consuming 200,000 L of lemon juice annually.

The British may have been slow to adopt Lind's findings, but they were faster at curing scurvy than their main naval opponents. An end to scurvy was one reason why the British, under the command of Admiral Lord Nelson, were able to maintain a sea blockade of France and ultimately win the 1805 Battle of Trafalgar against a larger force of scurvy-ridden French and Spanish ships. Nelson had clever tactics, but it helped that he did not have to fight while scurvy ravaged his crew.

So when you next find yourself looking up at Nelson's Column in Trafalgar Square, spare a thought for James Lind, who showed us that a curious researcher, conducting a simple randomised trial, really can change the course of history.

Acknowledgements

The author thanks the managing editor, John de New, two anonymous referees, Jon Baron, David Halpern, Frances Kitt, John Lavis, Deelan Maru and Eleanor Williams for insightful comments on earlier drafts.

Data Availability Statement

The author has nothing to report.

Endnotes

[1] Even within health, there are considerable differences. Among those papers published in top journals, 86% of studies of US medical interventions were randomised, compared with 18% of studies of US healthcare delivery interventions (Finkelstein and Taubman 2015a).

[2]Low-cost trials need not be low-impact trials. Jon Baron, president of the Coalition for Evidence-Based Policy, notes that most of the low-cost randomised trials that the Arnold Foundation funds are of more substantial interventions, and often include multi-year follow-up. Baron gives the example of a college advising program that cost US $4000 per student, had a large multi-site sample (2400 students) and a 6-year follow-up. Yet the research component of the study (not including the cost of program delivery, which would otherwise have been delivered in a non-randomised way) cost just US$200,000. The low cost was achieved by using a randomised lottery to allocate program slots (since the program was oversubscribed) and measuring all outcomes with administrative data such as college enrolment and completion.

References

Altmann, S., A. Falk, S. Jäger, and F. Zimmermann. 2018. "Learning About Job Search: A Field Experiment With Job Seekers in Germany." *Journal of Public Economics* 164: 33–49.

Australian Centre for Evaluation. 2025. *Government Administrative Data Sources for Evaluation in Australia*. Australian Treasury.

Banerjee, A. V. 2020. "Field Experiments and the Practice of Economics." *American Economic Review* 110, no. 7: 1937–1951.

Baron, J. 2018. "A Brief History of Evidence-Based Policy." *Annals of the American Academy of Political and Social Science* 678, no. 1: 40–50.

Behavioural Economics Team of the Australian Government (BETA). 2018. *Nudge vs Superbugs: A Behavioural Economics Trial to Reduce the Overprescribing of Antibiotics*. Department of the Prime Minister and Cabinet.

Berecochea, J. E., and D. R. Jaman. 1981. *Time Served in Prison and Parole Outcome: An Experimental Study: Report, No. 2*. Research Division, California Department of Corrections, Sacramento, CA.

Bergquist, L. F., B. Faber, T. Fally, M. Hoelzlein, E. Miguel, and A. Rodríguez-Clare. 2022. "Scaling Agricultural Policy Interventions." NBER Working Paper 30704, National Bureau of Economic Research, Cambridge, MA.

Bertrand, M., S. Djankov, R. Hanna, and S. Mullainathan. 2007. "Obtaining a Driver's License in India: An Experimental Approach to Studying Corruption." *Quarterly Journal of Economics* 122, no. 4: 1639–1676.

Bettinger, E. P., B. T. Long, P. Oreopoulos, and L. Sanbonmatsu. 2012. "The Role of Application Assistance and Information in College Decisions: Results From the H&R Block FAFSA Experiment." *Quarterly Journal of Economics* 127, no. 3: 1205–1242.

Bowman, L., F. Weidinger, M. A. Albert, E. T. A. Fry, and F. J. Pinto, Clinical Trial Expert Group and ESC Patient Forum. 2023. "Randomized Trials Fit for the 21st Century." *Journal of the American College of Cardiology* 81, no. 12: 1205–1210.

Briscese, G., and J. A. List. 2024. "Toward an Understanding of the Political Economy of Using Field Experiments in Policymaking." NBER Working Paper 33239, National Bureau of Economic Research, Cambridge, MA.

Buera, F. J., J. P. Kaboski, and R. M. Townsend. 2023. "From Micro to Macro Development." *Journal of Economic Literature* 61, no. 2: 471–503.

Christian, B. 2012. "The A/B Test: Inside the Technology That's Changing the Rules of Business." *Wired*, April 25.

Coalition for Evidence-Based Policy. 2013. "Randomized Controlled Trials Commissioned by the Institute of Education Sciences Since 2002: How Many Found Positive Versus Weak or No Effects." July.

Collins, R., L. Bowman, M. Landray, and R. Peto. 2020. "The Magic of Randomization Versus the Myth of Real-World Evidence." *New England Journal of Medicine* 382, no. 7: 674–678.

DellaVigna, S., W. Kim, and E. Linos. 2024. "Bottlenecks for Evidence Adoption." *Journal of Political Economy* 132, no. 8: 2748–2789.

Finkelstein, A., and M. J. Notowidigdo. 2019. "Take-Up and Targeting: Experimental Evidence From SNAP." *Quarterly Journal of Economics* 134, no. 3: 1505–1556.

Finkelstein, A., and S. Taubman. 2015a. "Randomize Evaluations to Improve Health Care Delivery." *Science* 347, no. 6223: 720–722.

Finkelstein, A., and S. Taubman. 2015b. "Using Randomized Evaluations to Improve the Efficiency of US Healthcare Delivery." J-PAL North America, Cambridge, MA.

Fukuhara, S., Y. Kataoka, T. Aoki, J. Green, S. Shimizu, and N. Toyoda. 2024. "International Collaboration and Commercial Involvement in Randomized Controlled Trials From 10 Leading Countries, 1997 Through 2019." *Cureus* 16, no. 5: e61205.

Global Commission on Evidence to Address Societal Challenges. 2022. *The Evidence Commission Report: A Wake-Up Call and Path Forward for Decisionmakers, Evidence Intermediaries, and Impact-Oriented Evidence Producers*. McMaster Health Forum.

Gomes, B. 2008. "Search Experiments, Large and Small." Google Official Blog, August 26.

Gonzalez-Navarro, M., and C. Quintana-Domeque. 2016. "Paving Streets for the Poor: Experimental Analysis of Infrastructure Effects." *Review of Economics and Statistics* 98, no. 2: 254–267.

Gosnell, G. K., J. A. List, and R. Metcalfe. 2016. "A New Approach to an Age-Old Problem: Solving Externalities by Incenting Workers Directly." NBER Working Paper 22316, National Bureau of Economic Research, Cambridge, MA.

Halpern, D. 2015. *Inside the Nudge Unit: How Small Changes Can Make a Big Difference*. WH Allen.

Halpern, D., and D. Maru. 2024. *Global Evidence Report: A Blueprint for Better International Collaboration on Evidence*. Behavioural Insights Team, Nesta and Economic and Social Research Council.

Hay, M., D. W. Thomas, J. L. Craighead, C. Economides, and J. Rosenthal. 2014. "Clinical Development Success Rates for Investigational Drugs." *Nature Biotechnology* 32, no. 1: 40–51.

Haynes, L., O. Service, B. Goldacre, and D. Torgerson. 2012. *Test, Learn, Adapt: Developing Public Policy With Randomised Controlled Trials*. Behavioural Insights Team, Cabinet Office.

Heller, S., H. A. Pollack, R. Ander, and J. Ludwig. 2013. "Preventing Youth Violence and Dropout: A Randomized Field Experiment." NBER Working Paper 19014. National Bureau of Economic Research, Cambridge, MA.

HM Treasury. 2020. *Magenta Book: Central Government Guidance on Evaluation*. HM Treasury.

Holz, J. E., J. A. List, A. Zentner, M. Cardoza, and J. Zentner. 2020. "The $100 Million Nudge: Increasing Tax Compliance of Businesses and the Self-Employed Using a Natural Field Experiment." NBER Working Paper 27666. National Bureau of Economic Research, Cambridge, MA.

Institute for Social Science Research. 2021. *Try, Test and Learn Evaluation*, ISSR, University of Queensland, Brisbane.

Leigh, A. 2018. *Randomistas: How Radical Researchers Changed Our World*. Black Inc.

Leigh, A. 2025. "Randomised Trials: The Seventh Phase of Good Government." Address to OECD International Workshop on Rigorous Impact Evaluation Approaches Including Randomised Controlled Trials, 5 February, OECD, Paris.

List, J. A. 2024. "Field Experiments: Here Today Gone Tomorrow?" *American Economist* 69, no. 2: 214–234.

Manson, J. E., C. J. Crandall, J. E. Rossouw, et al. 2024. "The Women's Health Initiative Randomized Trials and Clinical Practice: A Review." *Journal of the American Medical Association* 331, no. 20: 1748–1760.

Mazerolle, L., S. Bennett, P. Martin, M. Newman, D. Cowan, and S. Williams. 2021. "Evidence-Based Policing in Australia and New

Zealand: Empowering Police to Drive the Reform Agenda." In *The Globalization of Evidence-Based Policing: Innovations in Bridging the Research-Practice Divide*, edited by E. L. Piza and B. C. Welsh, 136–150. Routledge.

Muralidharan, K., P. Niehaus, and S. Sukhtankar. Forthcoming. "Identity Verification Standards in Welfare Programs: Experimental Evidence From India." *Review of Economics and Statistics*.

Roosevelt, F. D. 1932. "The New Deal." Oglethorpe University Address, Atlanta, GA, May 22.

Sanders, M., and J. Breckon. 2023. *The What Works Centres: Lessons and Insights From an Evidence Movement*. Policy Press.

Sherman, L. W., and D. Weisburd. 1995. "General Deterrent Effects of Police Patrol in Crime Hot Spots: A Randomized Controlled Trial." *Justice Quarterly* 12: 625–648.

Stampfer, M. J., W. C. Willett, G. A. Colditz, B. Rosner, F. E. Speizer, and C. H. Hennekens. 1985. "A Prospective Study of Postmenopausal Estrogen Therapy and Coronary Heart Disease." *New England Journal of Medicine* 313, no. 17: 1044–1049.

Taubman, S. L., H. L. Allen, B. J. Wright, K. Baicker, and A. N. Finkelstein. 2014. "Medicaid Increases Emergency-Department Use: Evidence From Oregon's Health Insurance Experiment." *Science* 343, no. 6168: 263–268.

Thodey, D., M. Carnegie, G. Davis, G. de Brouwer, B. Hutchinson, and A. Watkins. 2019. *Our Public Service, Our Future. Independent Review of the Australian Public Service*. Commonwealth of Australia.

Thomke, S. 2013. "Unlocking Innovation Through Business Experimentation." *European Business Review*, March 10.

Winzar, C., S. Tofts-Len, and E. Corpuz. 2023. *Disrupting Disadvantage 3: Finding What Works*. Committee for Economic Development of Australia.