THE AUSTRALIAN ECONOMIC REVIEW
THE UNIVERSITY OF MELBOURNE
MELBOURNE INSTITUTE
Check for updates
The Australian Economic Review, vol. 56, no. 4, pp. 431–441 DOI: 10.1111/1467-8462.12530

Evaluating Policy Impact: Working Out What Works

Andrew Leigh*

Abstract

Randomised trials frequently produce surprising findings, overturning conventional wisdom. During the twentieth century, randomised trials became commonplace within medicine, saving millions of lives. Randomised trials within government can now be conducted more cheaply, using administrative data. Just as it might be considered unethical to conduct a randomised trial if a program is indisputably effective, it might be considered unethical not to conduct a rigorous evaluation if a program lacks evidence. Developed within a robust ethical framework, and alongside community consultation, better evaluation can help governments save money and address social disadvantage.

1. Introduction

Social workers in schools always boost student outcomes. Drug offenders should not be treated differently. Malaria bed nets are more likely to be used if people pay for them. Seeing inside a jail will deter juvenile delinquents from becoming criminals.

All four statements sound perfectly sensible, don't they? Unfortunately, randomised trials suggest that all four are perfectly wrong. Let me explain.

In Britain, teachers, social workers and students all liked a pilot program placing social workers in schools. Then researchers at Cardiff and Oxford Universities ran a two-year randomised trial across 300 schools to test the program's impact. The results, reported this year, show no significant positive impact (Westlake et al. 2023). As a result, the planned national rollout has now been scrapped (Molloy 2023).

In New South Wales, a randomised trial of a specialised drug court shows that the tailored approach to drug offenders reduces recidivism (Lind et al. 2002). By treating their addiction, drug offenders' became much less likely to reoffend than if they had been sentenced through the traditional criminal justice system. Drug courts do not just help addicts—they also make the streets safer.

In Africa, some economists argued that free anti-malarial bed nets would not be valued by villagers, and might be used instead as makeshift fishing nets. So a randomised trial tested the take-up and use of free versus cheap bed nets. It turned out that free bed nets were far more popular, and equally likely to be used

* Parliament of Australia, Canberra, Australia. email: <andrew.leigh.mp@aph.gov.au>
This is an edited version of an address to the National Press Club in Canberra on 29 August 2023. Thanks to a range of experts, including Elisabeth Costa, David Halpern, Jon Lavis and officials in the Australian Centre for Evaluation for valuable feedback on earlier drafts, and to Ross Williams for his deft edits.

(Cohen and Dupas 2010). As the results of the randomised trials became clear, the World Health Organization switched its policy to favour free distribution of bed nets. The results of the experiment save thousands of lives every year.

In the United States, a policy known as 'Scared Straight' grew out of an Academy Award winning documentary. Juvenile offenders were brought into jails for a day, where they met hardened adult criminals. Low-quality evaluations—comparing those who took up the program with those who chose not to participate—suggested that it cut crime by up to half. But randomised trials told a different story, suggesting that participating in Scared Straight made youths substantially *more* likely to offend (Petrosino et al. 2013).

What these four examples have in common is that they used a randomised trial to evaluate the impact of a policy. Randomised trials have a long history in medicine, going back to James Lind's randomised trials of scurvy treatments in 1747, which helped save the lives of thousands of sailors. In the 1940s, randomised trials showed that antibiotics did not cure the common cold. In the 1950s, randomised trials showed that the polio vaccine was safe and effective.

Randomised trials helped drive the transformation from 'eminence-based medicine' to 'evidence-based medicine'. Until the end of the late nineteenth century, dangerous treatments such as bloodletting meant that doctors probably killed more patients than they saved. Even in the early twentieth century, Bayer was marketing heroin as a cough suppressant. The advent of randomised trials helped bring a what-works philosophy to medicine.

2. Every Good Evaluation Has a Good Counterfactual

Evaluations seek to answer a range of questions. Process evaluations can tell us whether a spending program was delivered on time, on budget and as intended. For new programs and small-scale pilots, it is also valuable to seek views from participants or service providers that could help address any weaknesses in the program design and implementation. After all, promising policies often under-perform due to poor implementation.

While these questions are important, process evaluations cannot tell us whether a program works, for whom, why, or in what circumstances. For that, we need an impact evaluation. And to determine what works, every impact evaluation—whether for medicine or policy—is trying to do one simple thing, that is, figure out the counterfactual. What would have happened if you did not take the pill, or did not participate in the program? This is what we get to see in the film *Sliding Doors*, when we follow Gwyneth Paltrow's two possible lives, according to whether she does—or does not—catch the train. It is what you got as a child when you re-read a *Choose Your Own Adventure* book.

In real life, we only get to see one version of reality, so we need to construct the alternative. Randomised trials do this by tossing a coin. Heads, you are in the treatment group. Tails, you are in the control group. Because luck determines whether you get the treatment, the two groups are equivalent at the outset. Thus any difference we see between them must be due to the intervention.

Low-quality evaluations sometimes construct the counterfactual by assuming that a person's outcomes would have remained unchanged in the absence of the intervention. This can give too much credit to the program. Most sick patients eventually get better. Most schoolchildren eventually become smarter. Most regions eventually grow. So any evaluation that assumes the world would otherwise have remained static is likely to produce a flawed result.

Jon Baron, who runs a US non-profit known as the Coalition for Evidence Based Policy, recently produced an example of this problem. Baron's example is based on results from the US Department of Health and Human Services' Comprehensive Child Development Program, which provides intensive case management services to low-income families with young children.

Over the 5 years of the program, employment rates for mothers in the program

doubled. This sounds impressive, until you realise that the study also had a randomly selected control group for which employment rates also doubled over the same period. As Baron notes 'If the Comprehensive Child Development Program had been evaluated in the usual non-rigorous way (examining employment outcomes without reference to a control group), it would've been deemed highly effective'. In reality, the program had no measurable effect on employment outcomes.

Problems also arise from evaluations that compare those who sign up for a program with those who do not. The kinds of parents who enrol their children in after-school tutoring are likely to be different from those who leave them to their own devices. The workers who choose job training are likely to be different from those who do not. In the early days of the COVID pandemic, a non-randomised study suggested that hydroxychloroquine was an effective treatment (Gautret et al. 2020). Subsequent randomised trials showed that it was not (Pathak et al. 2020).

This so-called 'selection effect' afflicts whole areas of social science. Thousands of studies have been published that compare health outcomes for people who choose to eat one kind of food instead of another. Increasingly, we are realising that this kind of study reveals a lot about the kinds of people who eat certain foods, but very little about the foods themselves.

Health writers Peter Attia and Bill Gifford point out that 'our food choices and eating habits are unfathomably complex', so observational studies are almost always 'hopelessly confounded' (Attia and Gifford 2023, 300). In other words, health studies based on comparing people who *choose* to eat different things may be as junky as a supersized burger with fries.

A better approach is that adopted by the US National Institutes of Health, which is conducting randomised nutrition studies. These require volunteers to live in a dormitory-style setting, where their diets are randomly changed from week to week. Nutritional randomised trials are costlier than nutritional epidemiology, but they have one big advantage: we can believe the results. They inform us about causal impacts, not mere correlations.

3. Changing the World is Harder than You Think

Rigorous evaluation is more likely to show up failure. A study published last year analysed ten different job training programs in the United States (Juras et al. 2022). Each program was evaluated in a sizeable randomised trial tracking earnings. After 6 years, only one program, YearUp, had a positive impact on earnings.

As the study points out, a lot needs to go right for a training program to boost earnings. It must have a sufficient impact on the credentials earned, those credentials must have labour market value, and the participants must find jobs. Training programs can fail because participants do not complete their studies, because the credentials have low economic returns, or because participants do not move into employment. The YearUp program produced a substantial wage return (around US$7000 a year), suggesting that it is possible to thread the needle. But nine out of ten programs did not perform on this outcome.

In education, the Coalition for Evidence-Based Policy analysed the randomised trials commissioned by the US Institute of Education Sciences and found that just one in 10 produced positive effects (Coalition for Evidence-Based Policy, 2013).

The problem is not confined to the public sector. The area where randomised trials are now most accepted is in the evaluation of new pharmaceuticals. In most advanced nations, getting public funding for a new drug requires that it go through Phase I, II and III clinical trials. Only one in ten drugs that look promising in the laboratory make it through all three phases and onto the market (Hay et al. 2014).

Another example from the business sector is Google, which—like many other successful companies—is constantly conducting randomised trials. Google estimates that just

one-fifth of these randomised trials help them improve the product (Thomke 2013).

These findings illustrate Rossi's Law, coined by sociologist Peter Rossi, which states that 'the better designed the impact assessment of a social program, the more likely is the resulting estimate of net impact to be zero' (Rossi 1987). This is not because high quality evaluation is cruel—it is because it is telling us the truth—designing programs that work better than what exists today is hard.

The people looking to improve education, medicine and technology platforms are smart, thoughtful and hardworking. They have access to a huge body of literature and oceans of data. When they produce a new intervention, they are probably confident that it works, and tempted to put it straight into the market. The fact that failure is more common than success does not suggest that program designers are foolish or careless, but that they are grappling with problems that are very difficult.

In the face of major challenges, low-quality evaluation is a hinderance, not a help. Using dubious impact evaluation techniques is like doing your running training with a slow watch. It might make you feel like you are fleet-footed, but when it comes to race day, you will eventually be shown up. That is why researchers in areas such as pharmaceutical development are committed to using randomised trials. They recognise the importance of accurately evaluating new treatments. They know that poor evaluation of medical treatments can cost money and lives.

In the face of difficult problems, we must bring more than a crash-or-crash-through mentality. We need to show up with a willingness to rigorously evaluate those solutions. We need to bring enough modesty to the task to acknowledge that answers that sound right may not always work in the real world. To generate and sustain a culture of continual learning, we need to be open to being proven wrong, and to use that information to do better the next time. We need to accept honest feedback—not pretend to get by with a dodgy wristwatch.

4. The Australian Centre for Evaluation

In the 2023 budget, the Australian Government announced the creation of the Australian Centre for Evaluation. Beginning with an annual budget of around $2 million, the centre's 14 staff members will work across the Australian Government and beyond to improve evaluation capabilities, practices and culture. A core role for the centre will be to champion high-quality impact evaluations, such as randomised policy trials.

Past reports have clearly shown the need to improve the quality of evaluation across government. Work done for the Thodey Review of the public sector found that the quality of evaluation was 'piecemeal'. Some high-quality evaluations have been conducted, including by the Behavioural Economics Team in the Department of Prime Minister and Cabinet. But in many other areas, the capacity to conduct rigorous evaluation is lacking.

The Australian Government already spends a considerable amount of money on evaluation. A report from the Australian Evaluation Society estimates that in 2021–22, the Commonwealth procured 224 evaluations from external consultants, at a total cost of $52 million (Australian Evaluation Society 2023). Because not all commissioned evaluations can be identified, the true volume of external evaluations commissioned from consultants by the Commonwealth may be larger still.

One problem for consultants is that there is not much incentive to undertake a high-quality evaluation. If Rossi's Law is right, then the better that consultants design their evaluation, the less likely they are to produce a report that shows the program worked. Which may make it harder for them to win the next contract.

That is why the Australian Government is also encouraging agencies to rebuild their own in-house evaluation capabilities and consider partnering with the Australian Centre for Evaluation to carry out high-quality evaluations (an approach consistent with the in-house consulting model being

created within the Department of Prime Minister and Cabinet).

Another reason that consultants' evaluations may fall short is if they are commissioned to produce evaluations late in the process, when it is difficult to identify a credible counterfactual. So the Australian Centre for Evaluation will also be working with government agencies to strengthen evaluation planning, especially in new budget proposals, and ensure that evaluation is considered at all stages of policy and not seen as an afterthought.

While the Australian Centre for Evaluation will operate across government, the Australian Government will not be compelling agencies to participate in evaluation partnerships. A high-quality evaluation is not like an audit, which can be conducted after the program has been rolled out. Good evaluation needs to be built into program design from the outset, which means working collaboratively with the departments deploying the programs. The Australian Centre for Evaluation will be complementing high-quality impact evaluations with other culturally safe evaluation methods that help gain an understanding of the lived experiences of Australians and support the delivery of better services.

During 2023, I have met with many of my ministerial colleagues to discuss which programs might be suitable for evaluation, and how we can drive higher standards of evidence to support decision making. Similar discussions are taking place at a public service level between the Australian Centre for Evaluation and other agencies, including the BETA team in the Department of Prime Minister and Cabinet.

We are working to change the old stereotype of randomised trials as slow and expensive, while ensuring relevant ethical, cultural and privacy considerations are at the centre of our thinking. It is true that blockbuster randomised trials such as Perry Preschool and the RAND Health Insurance Experiment took many years and cost many millions. But it is also possible to do things much more simply.

5. Tossing a Coin—Cheaply

We are not the first to think about the power of quick, simple policy trials to identify what works. In 2014, President Obama convened a White House conference on low-cost randomised trials. The result was a competition funded by the Laura and John Arnold Foundation that called for proposals to conduct simple, economical randomised trials, costing between US$100,000 and US $300,000. One trial provided support to low-income, first-generation students to enrol in college. Because it used administrative data, the evaluation cost just US$159,000 (and found significant positive impacts). The next year, the Arnold Foundation announced that the low-cost randomised trial initiative would continue, but that this time every program that received high ratings from its expert review panel would be funded (Arnold Foundation 2015). There is an opportunity for a major Australian philanthropic foundation to do likewise.

In Britain, the Education Endowment Foundation has also conducted many low-cost randomised trials. An analysis of its first 119 randomised trials in education found that three quarters cost less than £1 million—including both the cost of the intervention and the cost of the evaluation (Ames and Wilson 2016). In the context of many education programs, this is a relatively modest sum.

An example of a simple trial is the evaluation of Britain's tutoring program. In 2021, when school closures kept many pupils at home, the UK Government massively expanded tutoring programs. However, a significant challenge was to get disadvantaged pupils to attend tutoring. So the Education Endowment Foundation carried out small-scale randomised trials of three strategies to boost attendance. In the first intervention, pupils were sent reminder emails shortly before their sessions, so they did not accidentally forget. The second intervention gave tutors and pupils a 5-minute quiz about their hobbies, and emailed them afterwards to let them know what they had in common. The

third intervention gave tutors training in how to build a stronger relationship with their pupils. The only intervention that increased attendance was the hobby quiz, but it did so by a significant amount—around 7 per cent (Tagliaferri et al. 2022). There was good theory behind all three strategies. In practice, only one had the desired effect. That is why rigorous evaluation matters. And by making it economical, it can be built into the ordinary activities of government.

There is an old saying that if you think education is expensive, try ignorance. Likewise, if you think that evaluation is costly, try financing ineffective programs. Recall the example about social workers in schools, which was liked by everyone, but did not actually improve outcomes (Westlake et al. 2023). Prior to the trial, the UK Government had planned a national rollout of the program, but this was scrapped after the evaluation. This allowed money to be directed to more effective interventions. David Halpern, the head of the UK Behavioural Insights Team, informs me that it saved taxpayers around £1 billion a year—or enough to pay for the centre that generated it for the next 100 years.

Some randomised trials can be virtually free. If agencies are sending out letters, emails or text messages, then it should not cost much more to send two versions and see which works best. In business, this approach is known as A/B testing. In many firms, testing your ideas is just part of the corporate culture. Indeed, there are companies in which failing to have a control group will get you fired (I hasten to add that no public servant is going to lose their job because they did not have a control group). But the Australian Government is looking at how computer systems can be designed so that they make it easy for public servants to do low-fuss A/B tests—ensuring that government communications improve over time while saving time and money.

Another easy way that randomised trials can be used in government is to incorporate randomisation into the rollout of programs. During the global financial crisis, when the Australian Government decided to distribute households bonus payments to support the economy, it was clear that it would not be possible to have the money land in everyone's bank accounts on the same day. So a decision was made to take a list of all Australian postcodes, randomise the order, and have the payment schedule determined by that list. This had two advantages. It was fair, avoiding the need for the government to choose who would get the money first. And the resulting randomised trial allowed researchers to subsequently evaluate the short-term impact of the bonus payments on spending patterns (Aisbett, Brueckner and Steinhauser 2013).

6. Strong Democracies Use Rigorous Evaluation

The issue of fairness arises frequently in discussions of randomised trials. The most common criticism is that when we have an effective program, it is unethical to put people in the control group. A survey of Australian politicians found that half thought randomisation was unfair (Ames and Wilson 2016). Where the evidence is solid, I agree. However, I would also argue that if we do not know whether the program works, it is unethical *not* to conduct a rigorous evaluation if one is practically feasible.

Any discussion of ethics also needs to bear in mind the possibility that programs are harming the people they are intended to help. In the 1990s, the US Congress established a program called 21st Century Community Learning Centers, which provided US$1 billion in funding each year to high-poverty schools, to provide after-school activities, such as homework assistance, as well as activities such as basketball. When asked by researchers to identify the impact of the centres, teachers said that students who attended the centres had made improvement over the year in their academic performance, motivation, attentiveness and classroom behaviour (Naftzger et al. 2006).

Then the government commissioned a randomised trial of the program, using the fact that centres were oversubscribed to

conduct a lottery for the slots available. Comparing lottery winners and losers, the randomised trial found that those who won a spot in the after-school centres did no better academically. But on behavioural outcomes, they did considerably worse (James-Burdumy et al. 2008). The rate of school suspensions among attendees was around 50 per cent higher. Subsequent studies pointed to a modelling effect—that delinquent boys were encouraging each other to act up. As the father of three boys, I am no stranger to this effect, but I would not have predicted it of the learning centres. Nonetheless, it is a fact that for those in the randomised trial, it was better to be in the control group than the treatment group.

Carrying out randomised trials can also help strengthen our democracy (Tanasoca and Leigh 2023). By building a strong evidence base for programs, citizens can see that government is crafting programs based on what works, rather than blind ideology or partisan self-interest. It is no coincidence that authoritarian regimes have been the most resistant to science and evidence. Building a better feedback loop demonstrates to the public that the focus of government is on practical problem solving. And because the results of randomised trials are intuitively easy to understand, they bring the public into the discussion, allowing everyone to see what works.

7. The Credibility Revolution

All this is taking place against the backdrop of a 'credibility revolution' in the social sciences. A 2015 study led by psychologist Brian Nosek looked at 100 studies published in top journals, and found that only one in three could be replicated (Open Science Collaboration 2015). Where the results could be replicated, the size of the effect shrank to about half of that found by the original study. This may be due in part to a tendency by leading academic journals to publish exciting results. If results vary by chance, then the best journals will end up publishing inflated results (Ioannidis 2005).

Alas, there is more than chance at work. Over recent years, a number of top researchers, including food researcher Brian Wansink, behavioural scientist Francesca Gino and psychologist Dan Ariely, have been accused of fabricating data for their studies. Ironically, a study of dishonesty among car dealers appears to have been based on fake data. A slew of studies have been withdrawn. In one clue as to the size of the problem, anonymous surveys find that 2 per cent of social scientists admit to falsifying data (Fanelli 2009).

Whether the problem arises from luck or fraud, the best antidote to dodgy research is good research. A decade ago, when a leading psychology journal published work by Daryl Bem purporting to prove extrasensory perception (Bem 2011), other researchers were quick to carry out replication studies that found no such effects (Ritchie et al. 2012). This is the way science should advance—with other researchers rigorously testing surprising findings to see whether they hold up.

Randomistas within government have been at the forefront of these efforts. Indeed, one of the first to discover that Dan Ariely's signature studies did not replicate was Ariella Kristal, then a researcher at the UK Behavioural Insights Team (Kristal et al. 2020). Results from randomised trials conducted by government researchers tend to be smaller in magnitude than those from randomised trials run by academics; most likely because the government randomistas are using larger samples (Della Vigna and Linos 2022). In other words, randomised trials in government do not just help improve policies—they have also helped to rein in some of the wilder claims in the published literature.

8. Big Data, Better Results

Naturally, big data creates big responsibilities. With government holding more personal data than ever before, this raises the importance of keeping Australians' data safe, of maintaining proper privacy protections, and of using these data wisely.

Aggregated to preserve anonymity, administrative data can be used to help government work better. As Chief Statistician David Gruen puts it, these data are 'becoming increasingly important to provide the evidence base for policy, community-level insights, and program evaluation capability' (Gruen 2023).

Most people do not enjoy doing surveys, and by using administrative data, governments can avoid expensive surveys with their declining response rates. This means that governments can include everyone in the research, rather than having to focus the study only on the subset of people who choose to complete the survey.

Recall the randomised trials of the NSW Drug Court? Its results were based on administrative data in which participants' reoffending was determined based on the names of those who came before the court again. Or take a study carried out nationally in 2016–2017, which sent gentle letters to doctors with the highest rates of antibiotic prescriptions, pointing out that they were among the top third of 'superprescribers' in their region, and reminding them of the need to reduce antimicrobial resistance (Australian Government 2018). To find out whether the intervention worked, the researchers used administrative data on prescribing rates. No survey required.

As an added bonus, studies that have both administrative measures and survey measures have found that in areas such as hospital visits, administrative measures are more accurate (Taubman et al. 2014).

9. Conclusion

In 2021 and 2022, I served as one of 25 commissioners on the *Global Commission on Evidence to Address Societal Challenges*. Led by John Lavis and a secretariat at McMaster University in Canada, the commissioners were drawn from across the globe. We concluded that

> Evidence … is not being systematically used by government policymakers, organizational leaders, professionals and citizens to equitably address societal challenges. Instead decision-makers too often rely on inefficient (and sometimes harmful) informal feedback systems. The result is poor decisions that lead to failures to improve lives, avoidable harm to citizens, and wasted resources.

Our *Global Commission on Evidence* report (Global Commission on Evidence to Address Societal Challenges 2022) proposed that the World Bank devote an upcoming World Development Report to evaluation, that national governments review their use of evidence, and that budgeting take account of evidence. We also suggested that citizens better use evidence, making decisions on their wellbeing based on the best evidence, choosing products and services that are backed by evidence, and donating to causes that are evidence-based.

For my own part, evidence has shaped how I live my life (Leigh 2018). Randomised trials of daily vitamin supplements persuades me that they do not have much benefit for otherwise healthy people. As a donor, rigorous evidence from GiveWell.org has persuaded me to donate to their top-ranked causes. As a runner, randomised trials of marathoners has convinced me that compression socks speed recovery. If the evidence changes, I am open to changing my diet, my donations and even my socks.

Underpinning the philosophy of randomised trials is a curiosity about the world, a willingness to experiment and a modesty about our knowledge. Many of the problems that we face in public policy are hard. If it was easy to close life expectancy gaps, educational gaps or employment gaps, then past generations would have done it already. The fact that these challenges persist means that good intentions are not enough.

In the decades since randomised trials became broadly accepted as the best way of evaluating medical treatments, millions of lives have been saved. From childhood leukaemia to heart attacks, survival rates have improved dramatically and continue to improve. That is not because every treatment emerging from the laboratory has worked. It is because medicine has subjected those treatments to rigorous evaluation.

The Australian Centre for Evaluation seeks to take the same approach to policy—testing new ideas with the same methods we use to test new pharmaceuticals. We are looking to make rigorous evaluation a normal part of government: from A/B testing the wording of government letters to using administrative data to evaluate new initiatives.

Public servants can become better *consumers* of evidence. When claims are made about the effectiveness of a program, ask about the quality of that evidence. Is it a single before–after study, or a systematic review of multiple randomised trials? Each of us can work to raise the evidence bar.

Public servants can also help *produce* evidence about what works. If agencies spot an opportunity to run a high-quality evaluation, we encourage them to engage with the Australian Centre for Evaluation.

Much of the expertise on randomised trials already exists in academia. We hope that the Australian Centre for Evaluation can strengthen partnerships between government agencies and academic experts who are already conducting rigorous evaluations. The Australian Government also hopes to partner with state and territory governments, non-profits and philanthropic foundations to improve the quality of evaluation nationwide.

Over time, embedding evaluation in the work of governments could take many forms. Rather than running pilot studies, governments might ensure that all small-scale trials have a credible control group. When policies are rolled out to different sites over time, governments could consider building randomisation into the rollout, guaranteeing a rigorous evaluation. When programs are oversubscribed, governments might use a lottery approach to allocate the scarce places, and follow up the outcomes of both groups. When the Australian Government is distributing funds to states and territories, it would be possible to provide resources for those jurisdictions that are willing to conduct rigorous experiments. When allocating resources to non-profit organisations, governments could potentially provide more support to those with programs backed by the best evidence.

Finally, everyone who cares about opportunity should support the mission to conduct more randomised trials and improve the quality of evaluation. When government fails, the most affluent have private options—private transport, private education, private healthcare and private security. It is the poorest who rely most on government, and the most vulnerable who stand to gain when government works better. Disadvantaged Australians do not need ideology, they need practical solutions that improve their lives. Better evaluation would not just boost the productivity of government; it can also shape a more equal nation.

References

Aisbett, E., Brueckner, M. and Steinhauser, R. 2013, '*Fiscal Stimulus and Households' non-durable consumption expenditures: Evidence from the 2009 Australian Nation Building and Jobs Plan, Crawford School Research*'. Canberra, Australian National University.

Ames, P. and Wilson, J. 2016, 'Unleashing the potential of randomised controlled trials in Australian governments, M-RCBG Associate Working Paper No. 55, Harvard Kennedy School, Cambridge, MA'.

Arnold Foundation. 2015, 'Arnold Foundation announces expanded funding for low-cost randomized controlled trials to drive effective social spending', Laura and John Arnold Foundation, Washington DC'.

Attia, P. and Gifford, B. 2023, 'Outlive: The science and art of longevity'. Harmony, New York.

Australian Evaluation Society. 2023, '*The state of evaluation in Australia: A study of current evaluation volume, drivers, approaches, and trends in Australia*', Australian Evaluation Society, Carlton South, Vic.

Australian Government. 2018, '*Nudge vs Superbugs: A report into a behavioural economics trial to reduce the overprescribing of antibiotics*'. Canberra, Department of Health and Department of Prime Minister and Cabinet.

Bem, D. J. 2011, 'Feeling the future: Experimental evidence for anomalous retroactive influences on cognition and affect', *Journal of Personality and Social Psychology*, vol. 100, no. 3, pp. 407–425.

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'.

Cohen, J. and Dupas, P. 2010, 'Free distribution or cost-sharing? Evidence from a randomized malaria prevention experiment', *Quarterly Journal of Economics*, vol. 125, no. 1, pp. 1–45.

Della Vigna, S. and Linos, E. 2022, 'RCTs to scale: Comprehensive evidence from two nudge units', *Econometrica*, vol. 90, no. 1, pp. 81–116.

Fanelli, D. 2009, 'How many scientists fabricate and falsify research? A systematic review and meta-analysis of survey data', *PLoS ONE*, vol. 4, no. 5, pp. e5738.

Gautret, P., Lagier, J. C., Parola, P., Meddeb, L., Mailhe, M., Doudier, B., Courjon, J., Giordanengo, V., Vieira, V. E., Dupont, H. T. and Honoré, S. 2020, 'Hydroxychloroquine and azithromycin as a treatment of COVID-19: Results of an open-label non-randomized clinical trial', *International journal of antimicrobial agents*, vol. 56, no. 1, pp. 105949.

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*. Hamilton, McMaster Health Forum.

Gruen, D. 2023, 'Supporting analysis of the life course. Speech delivered 15 August 2023 Life Course Centre Data for Policy Summit, Canberra'.

Hay, M., David, W. T., Craighead, J. L., et al. 2014, 'Clinical development success rates for investigational drugs', *Nature biotechnology*, vol. 32, no. 1, pp. 40–51.

Ioannidis, J. P. 2005, 'Why most published research findings are false', *PLoS medicine*, vol. 2, no. 8, pp. e124.

James-Burdumy, S., Dynarski, M. and Deke, J. 2008, 'After-school program effects on behavior: Results from the 21st Century Community Learning Centers program national evaluation', *Economic Inquiry*, vol. 46, no. 1, pp. 13–18.

Juras, R., Gardner, K., Peck, L. and Buron, L. 2022, 'Summary and insights from the long-term follow up of ten PACE and Hpog 1.0 job training evaluations. In 2022 APPAM Fall Research Conference. APPAM'.

Kristal, A. S., Whillans, A. V., Bazerman, M. H., Gino, F., Shu, L. L., Mazar, N. and Ariely, D. 2020, 'Signing at the beginning versus at the end does not decrease dishonesty', *Proceedings of the National Academy of Sciences*, vol. 117, no. 13, pp. 7103–7107.

Leigh, A. 2018, '*Randomistas: How radical researchers are changing our world*'. New Haven, Yale Press.

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.

Molloy, D. 2023. 'Social workers in schools: Why we are not recommending investment', What Works for Children's Social Care, London.

Naftzger, N., Kaufman, S., Margolin, J. and Ali, A. 2006, '*21st Century Community Learning Centers (21st CCLC) Analytic Support for Evaluation and Program Monitoring: An Overview of the 21st CCLC Program: 2004–05*'. Naperville, IL: Learning Point Associates, Report prepared for the U.S. Department of Education.

Open Science Collaboration 2015, 'Estimating the reproducibility of psychological science', *Science*, vol. 349, no. 6251, pp. aac4716.

Pathak, S. K., Salunke, A. A., Thivari, P., Pandey, A., Nandy, K., Ratna, H. V., Pandey, S., Chawla, J., Mujawar, J., Dhanwate, A. and Menon, V. 2020, 'No benefit of hydroxychloroquine in COVID-19: Results of systematic review and meta-analysis of

randomized controlled trials', *Diabetes & Metabolic Syndrome: Clinical Research & Reviews*, vol. 14, no. 6, pp. 1673–1680.

Petrosino, A., Turpin-Petrosino, C., Hollis-Peel, M. E. and Lavenberg, J. G. 2013, 'Scared straight and other juvenile awareness programs for preventing juvenile delinquency: A systematic review', *Campbell Systematic Reviews*, vol. 9, no. 1, pp. 1–55.

Ritchie, S. J., Wiseman, R., French, C. C. and Gilbert, S. 2012, 'Failing the future: Three unsuccessful attempts to replicate Bem's 'Retroactive Facilitation of Recall' effect', *PLoS ONE*, vol. 7, no. 3, pp. e33423.

Rossi, P. H. 1987, 'The iron law of evaluation and other metallic rules', *Research in Social Problems and Public Policy*, vol. 4, pp. 3–20.

Tagliaferri, G., Chadeesingh, L., Xu, Y., Malik, R., Holt, M., Bohling, K., Sreshta, P. and Kelly, S. 2022. 'Leveraging pupil-tutor similarity to improve pupil attendance', Education Endowment Foundation and Behavioural Insights Team.

Tanasoca, A. and Leigh, A. 2023, 'The democratic virtues of randomized trials', Moral Philosophy and Politics.

Taubman, S. L., Allen, H. L., Wright, B. J., Baicker, K. and Finkelstein, A. N. 2014, 'Medicaid increases emergency-department use: Evidence from Oregon's Health Insurance Experiment', *Science*, vol. 343, no. 6168, pp. 263–268.

Thomke, S. 2013, 'Unlocking innovation through business experimentation', European Business Review.

Westlake, D., Pallmann, P., Lugg-Widger, F., White, J., Forrester, D., Petrou, S. and Daer, S. 2023. 'The Social Workers in Schools (SWIS) trial: An evaluation of school-based social work', What Works for Children's Social Care, London.