Australian Government
The Treasury

Australian Centre for Evaluation ace

Understanding and interpreting statistical results in impact evaluations

Dr Vy Nguyen, Dr Scott Copley, Harry Greenwell
Australian Centre for Evaluation (ACE), Commonwealth Treasury

evaluation.treasury.gov.au

September 2025

Acknowledgment of Country

Did you know eating chocolate helps you lose weight?

(Story of a prank went viral)

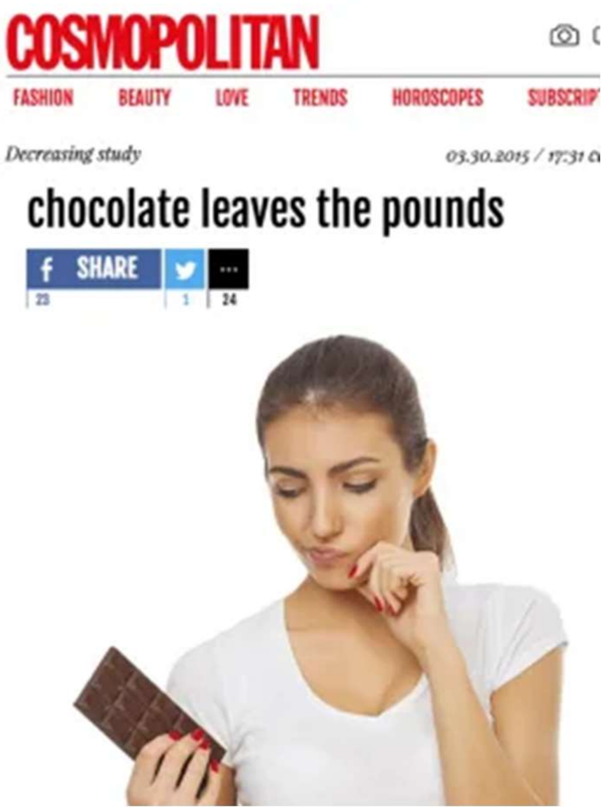
Chocolate weight-loss prank

15 participants

18 variables measured

Significant weight loss

(This is fake news)



Constructive scepticism is needed to evaluate causal evidence

Respect the evidence, but asking the right questions before we let a finding become a headline or, worse, a policy

Some starting caveats

Today we focus exclusively on quantitative assessments of impact, only on 'frequentist' approach to data analysis.

Research design, statistics and reporting matter – together they determine evidence quality and thus how much weight the evidence deserves

Our stance – respect the evidence, but also interrogate it

Adopt constructive scepticism: respect the work but ask critical questions

No perfect studies
All research has limitations and uncertainty

Context matters
The evidence of any study is context-specific

Systemic biases
The research itself has systematic issues

Outline: things to do when reading/judging a quant report

Assess Claims and Design

Determine if the study is descriptive or causal and if the design supports the claim.

Evaluate Sampling and Size

Understand who is included in the research and the number of participants.

Interpret Statistics

How to interpret statistical concepts like confidence intervals and p-values.

Identify Good vs Bad Practice

Recognise what to do and avoid in research.

Assessing claims in evaluation

Is a causal claim being made?

What type of claim is being made?

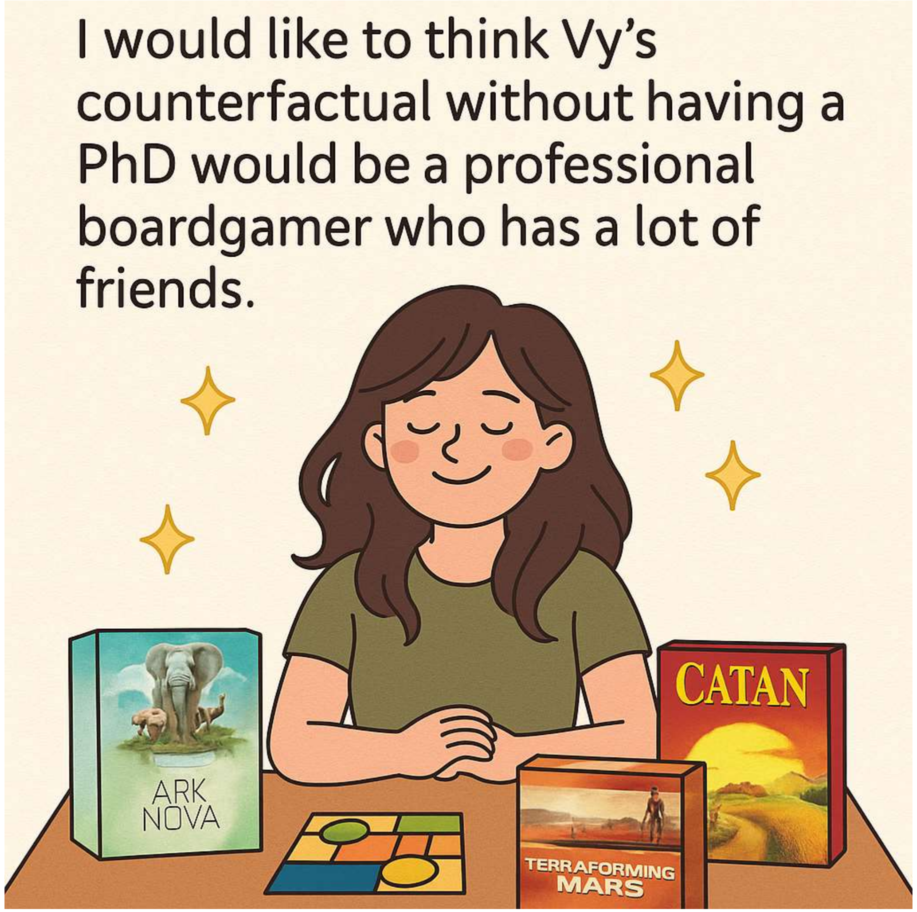
Descriptive Claim

Provides information about what happened but does not establish causation.

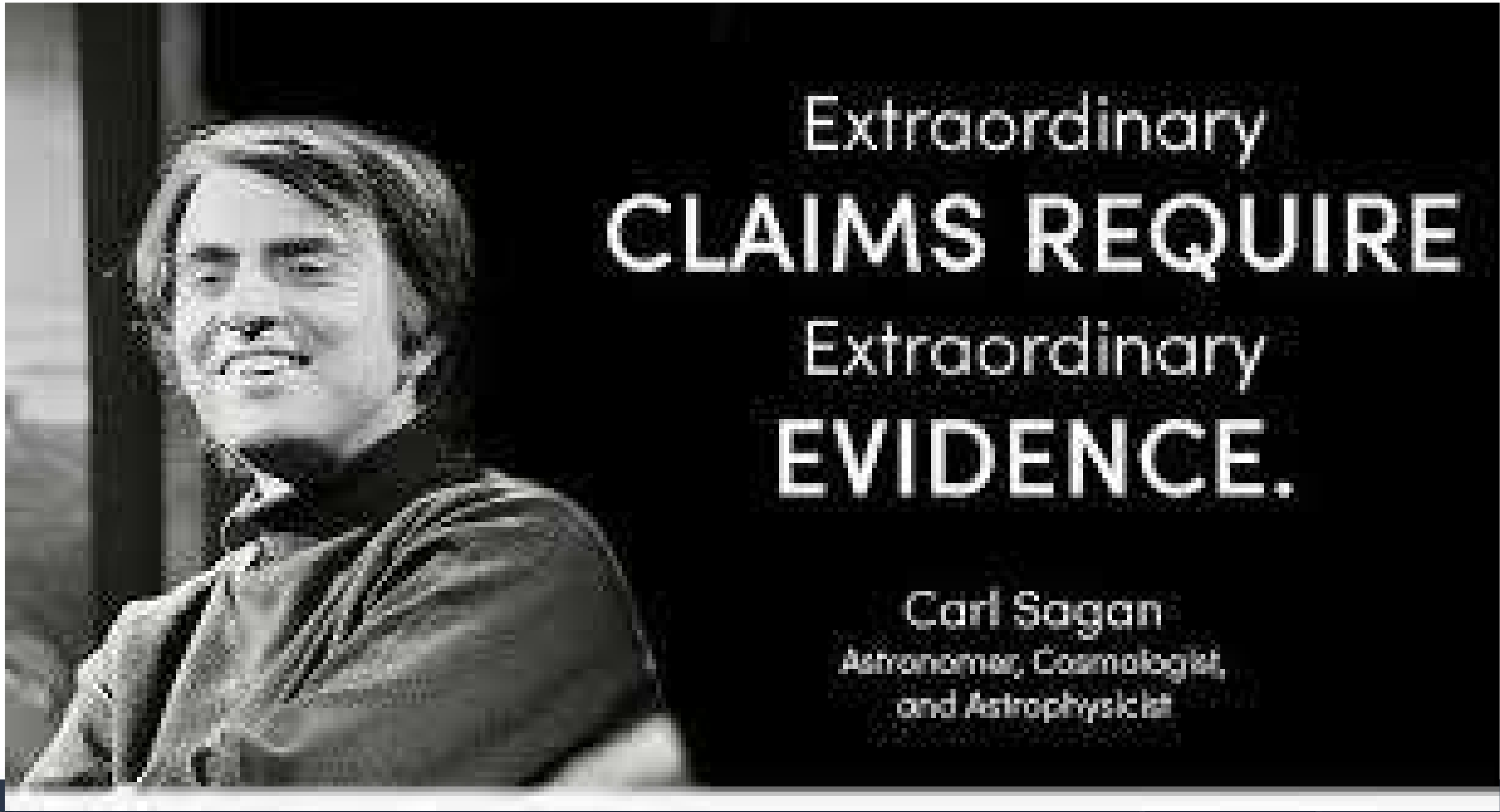
Causal Claim

Asserts that the program caused the change, requiring a counterfactual.

The counterfactual – what if the program never happens?



Is such causal claim feasible?



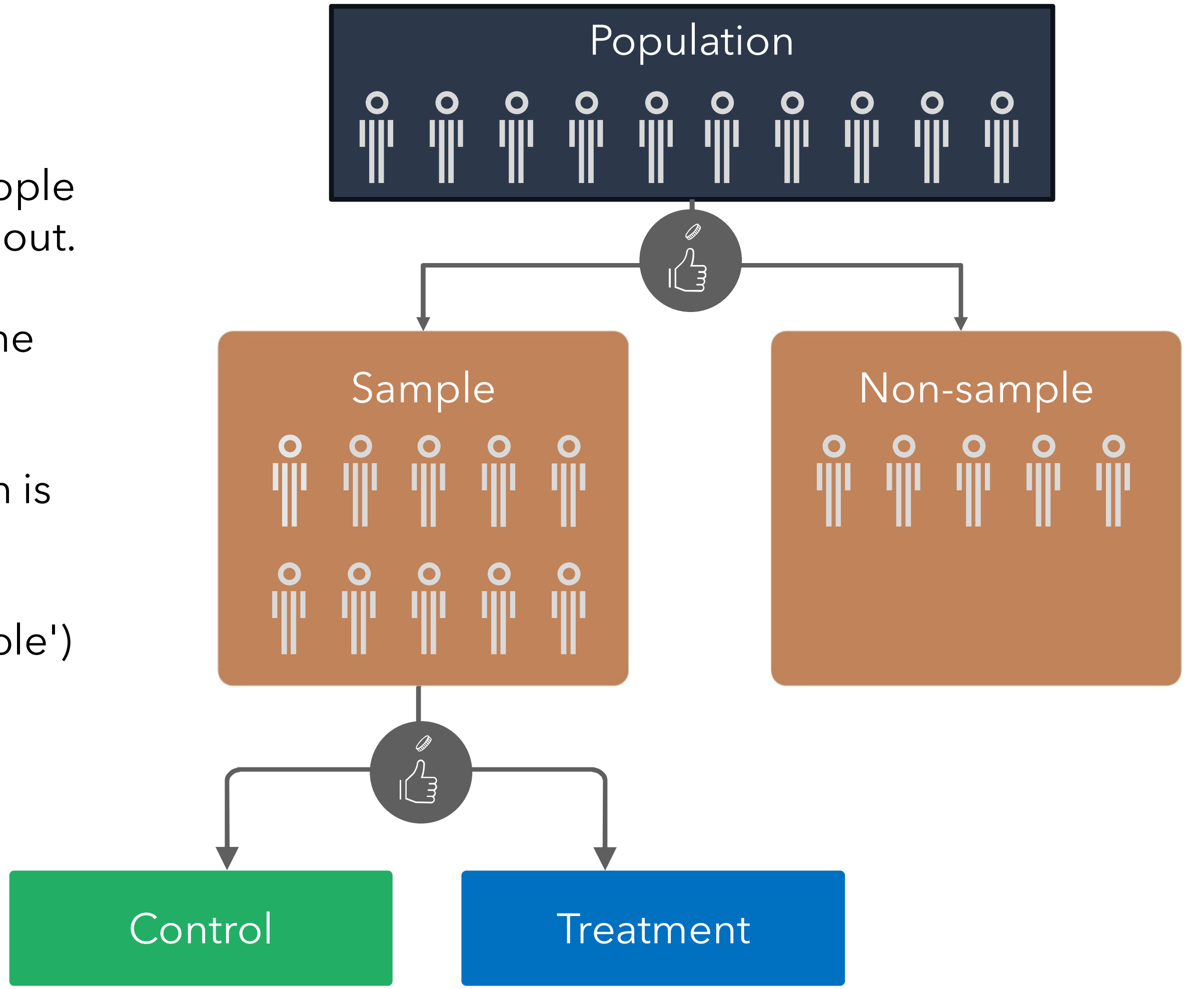
Can the design support the claim?

Experimental	Randomised Controlled Trials	Well-conducted RCTs can give the greatest confidence in causal claims
Design based (Quasi-experimental)	Found or 'natural' experiments Regression Discontinuity Design Difference-in-Differences Instrumental variables	Can generate robust causal evidence, if their assumptions hold
Control-on-observables (Non-experimental)	Matching Regression	Causal claims rely on a strong assumption that all confounders are controlled for

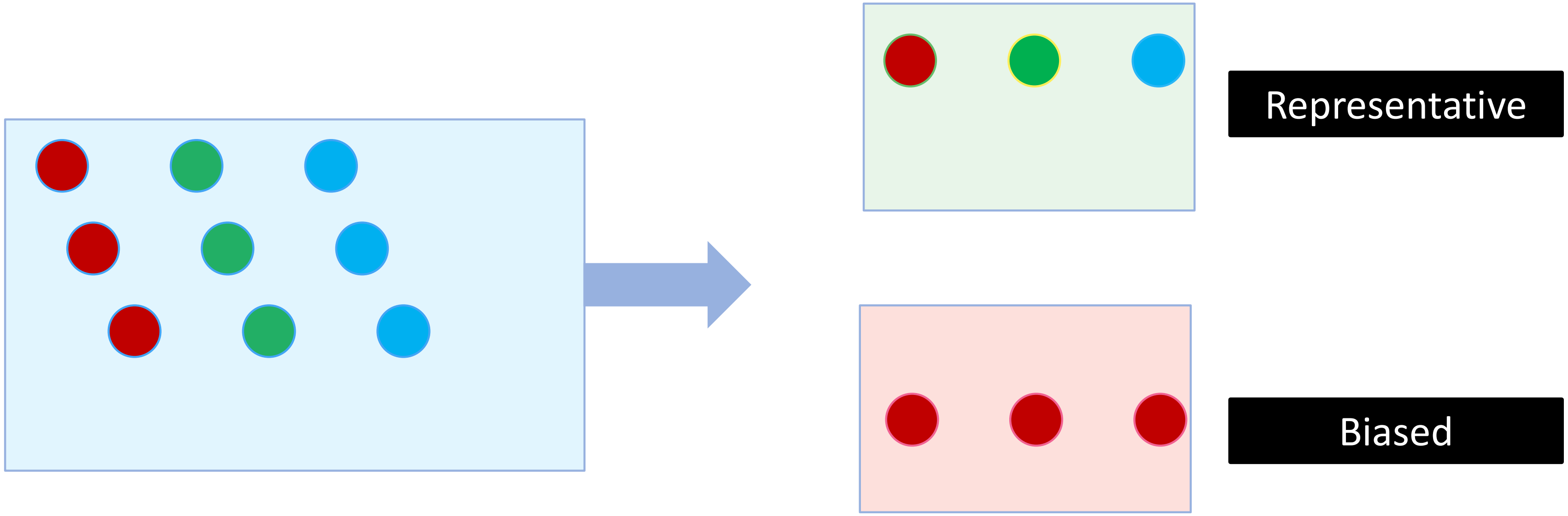
Does the research have the right sample?

Participants

- Need to carefully define the group of people (the population) that you wish to learn about.
- This whole population might be part of the RCT.
- If not, a random sample of the population is ideal.
- Eligibility criteria (or a 'convenience sample') are often necessary but impacts generalisability.



Sampling & bias



Does the research have enough people?

False positive versus false negative

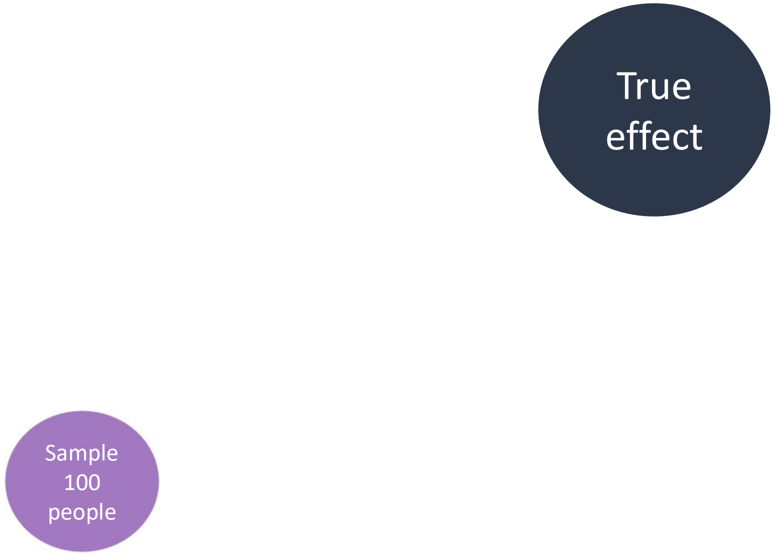
Type II error – false negative

Type I error : false positive

Power and sample size

Statistical power is the probability of detecting a true effect of a given size.

Power and sample size



Statistical power is the probability of detecting a true effect of a given size.

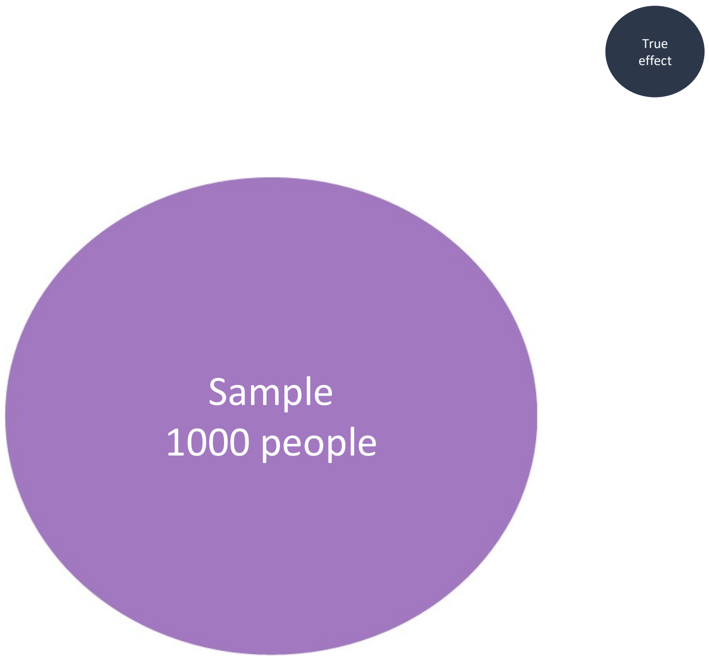
- A smaller sample size makes it harder to detect an effect, if there is one

Power and sample size

Statistical power is the probability of detecting a true effect of a given size.

- It is harder to detect a small effect size

Power and sample size



- A larger sample size is needed to detect smaller effects

Things to look out for:

Representativeness

Check if the sample resembles the target population.

Random sampling

Check if the sampling is random or convenient

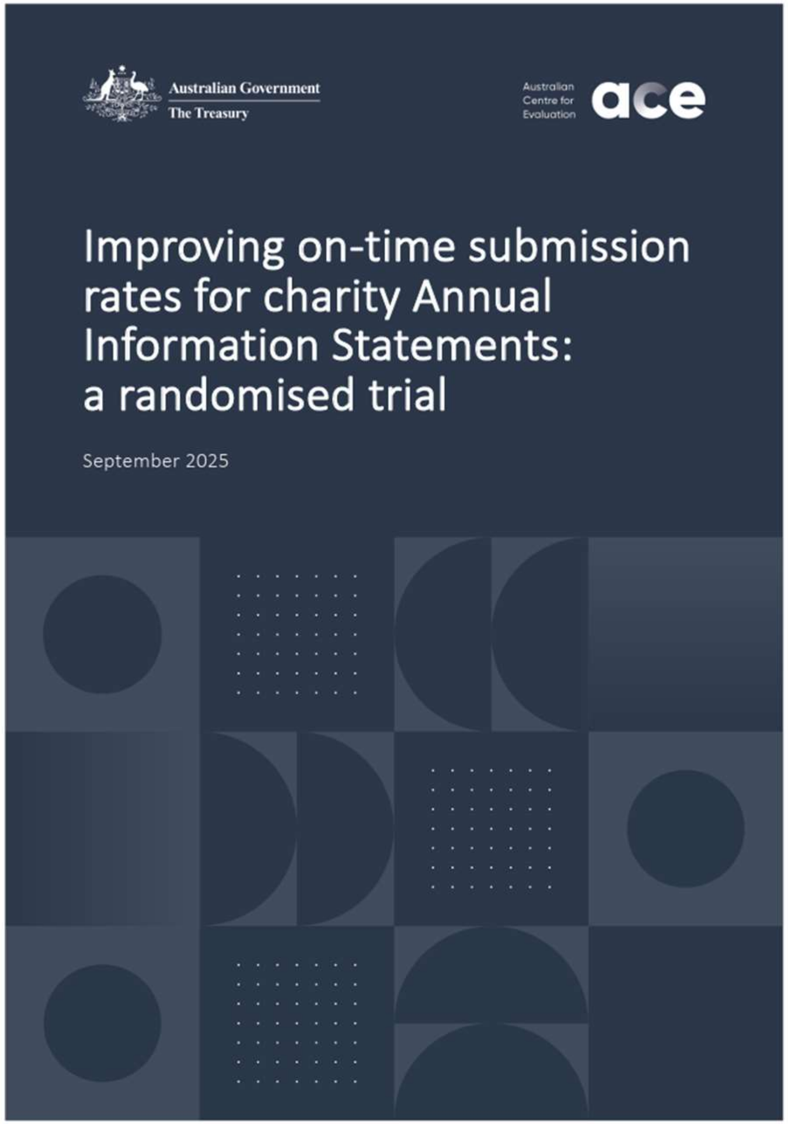
Statistical power

Ensure the sample size is large enough to detect a meaningful effect.

From average to effect size: what the stats really tell us

Recall the last evaluation you came across and how the findings were shown - maybe a percentage change, maybe a p-value, maybe just a headline that the program worked?

….. and if your mind goes blank, may we cheekily suggest checking out our fresh-off-the-press evaluation or check out our evaluation library?



Entirely suitable as bedtime reading.

Very rigorous
(admittedly self-proclaimed)

Available under the publications tab of our website: evaluation.treasury.gov.au

Statistical results

Effect size & ‘practical significance’

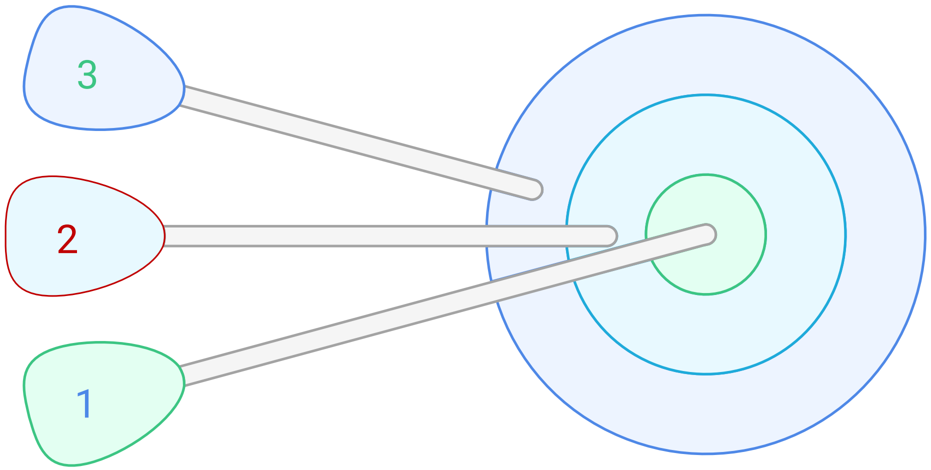
Compares program effects with others

Confidence interval

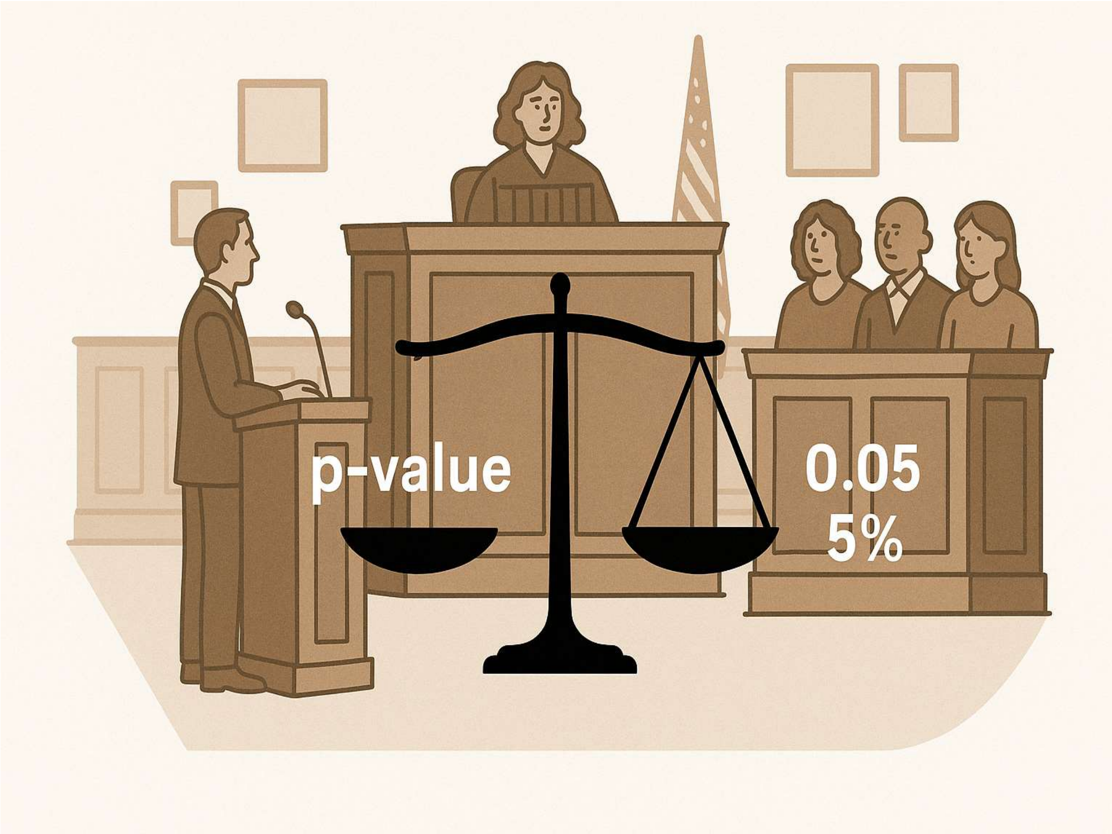
Measures the magnitude/range of magnitude of program effects

P-value & ‘statistical significance’

Determines program effectiveness?



p-value and level of ‘statistical significance’



p-value: If the defendant were truly innocent (the “null” hypothesis) , how surprising is the evidence we’ve just seen?

A small p-value = very surprising evidence if the null were true -> we lean toward rejecting innocence.

The threshold (often 0.05) is like our “beyond reasonable doubt” bar

p-values, aka, does the program work?

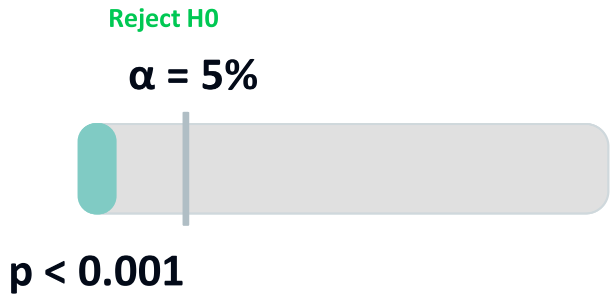
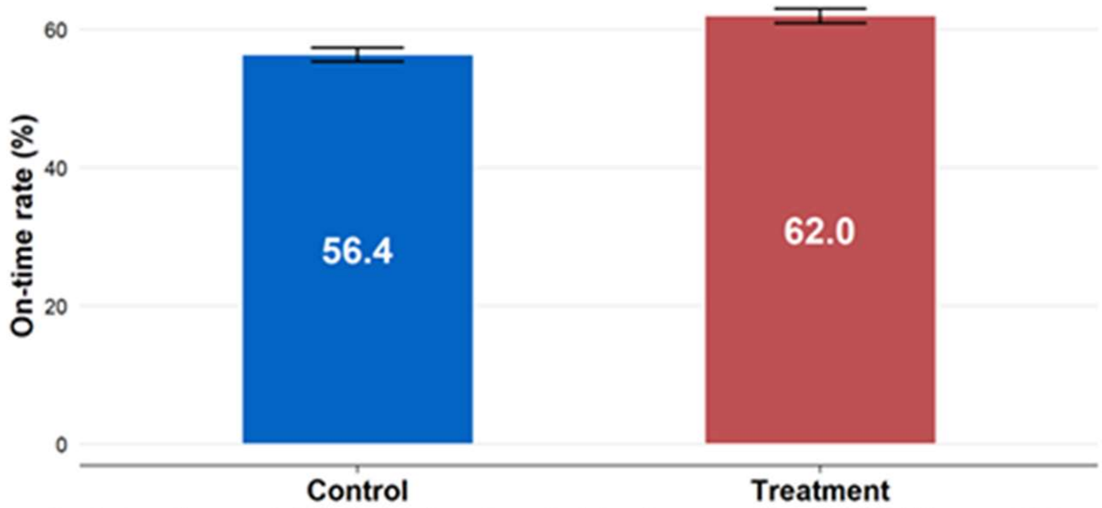
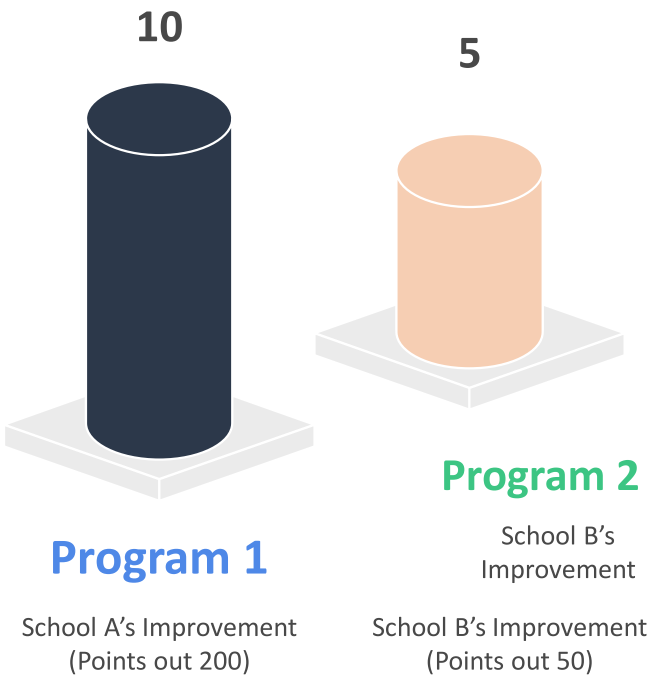


Figure 2. Change in on-time AIS submission rates



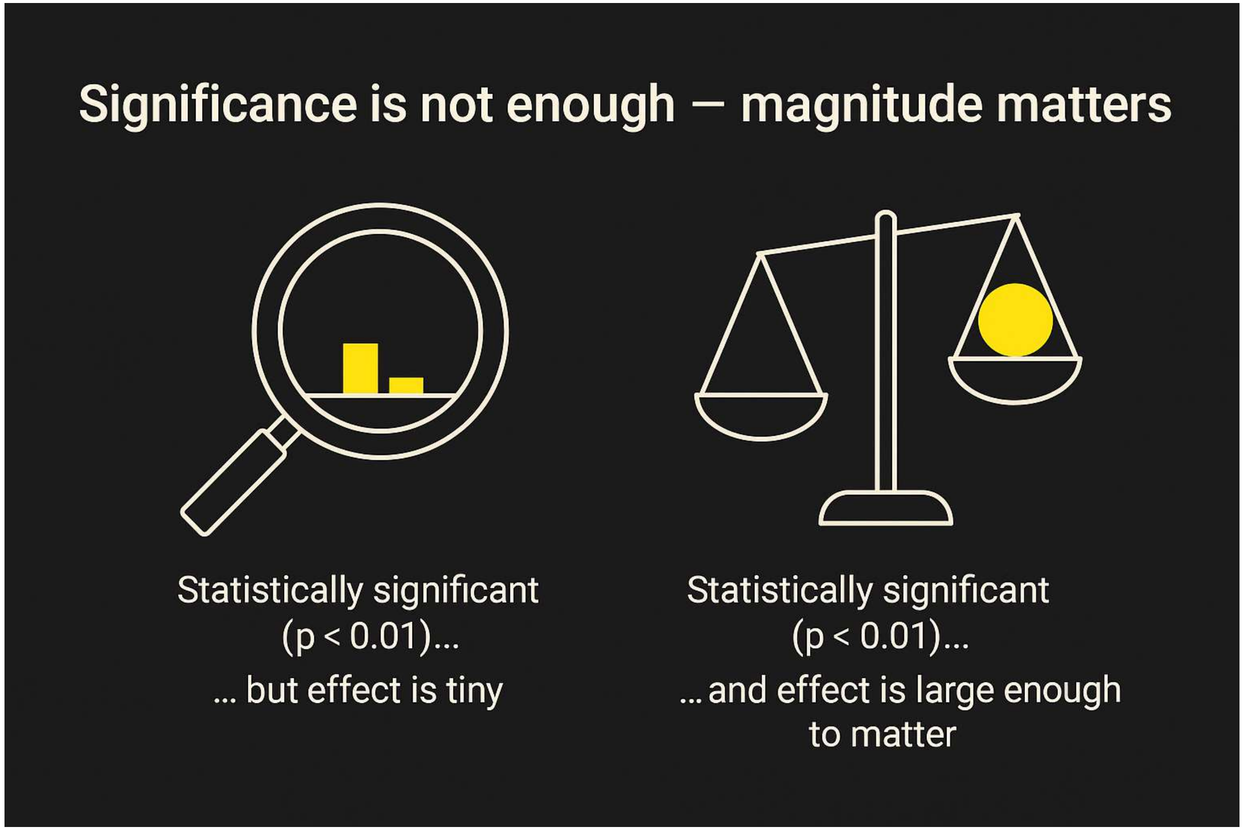
Note: Results from linear probability model of on-time submission, by treatment status. The difference of 5.6 percentage points was statistically significant ($p<0.001$). N=15,000.

Effect versus Standardised effects



- raw differences depend on the scale of measurement.
- A **standardised effect size** solves this by putting results on a common scale.
- So instead of saying "+5 points" or "+10 points", you say "the program improved scores by **0.5 standard deviations**"

Statistical versus practical significance

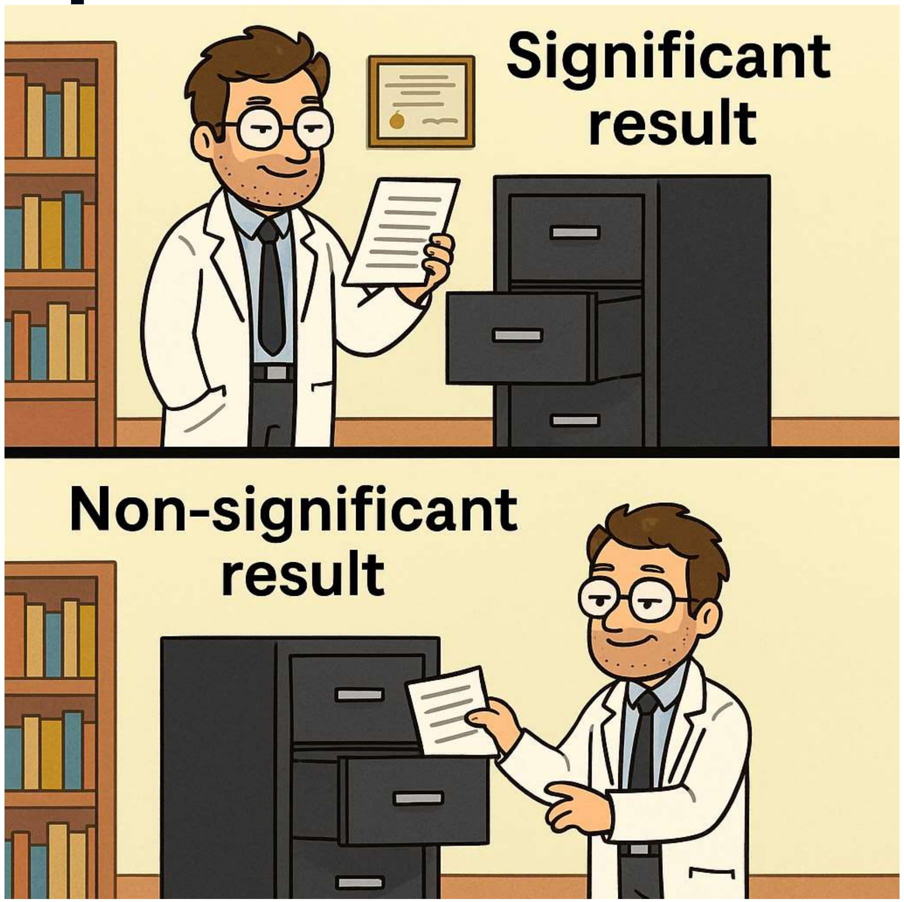


Did the researchers tie their hand?

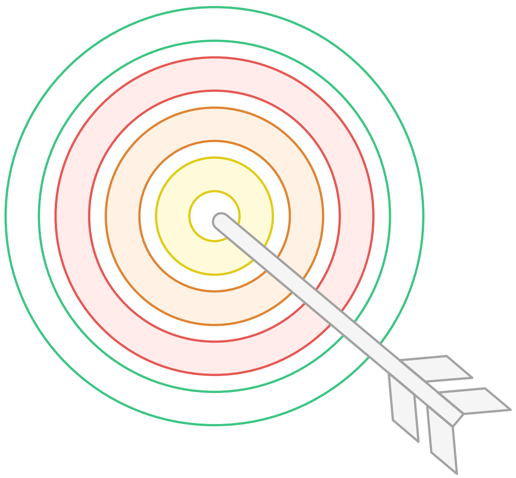
(A healthy self-roast on bad practice in quantitative research)

We have heard lots about publication bias in social science

A New England Journal of Medicine paper examined the publication rate of trials relating to antidepressants. (Turner et al. 2008) 36 of 37 (97%) antidepressant trials with positive results were published, compared with 3 of 22 (14%) antidepressant trials with negative results.



Pathways that make a finding look stronger:



Chance

Randomly achieving statistically significant results

Multiple Hypothesis/Comparisons

Increasing chances of false positives

Researcher flexibility and p-hacking

Exploring various analytical approaches until we get p-value <5% (another way to get a 'false positive')

Lack of transparency in reporting

Methods and negative results may be hidden, making findings look more robust than they are.

Multiple hypothesis

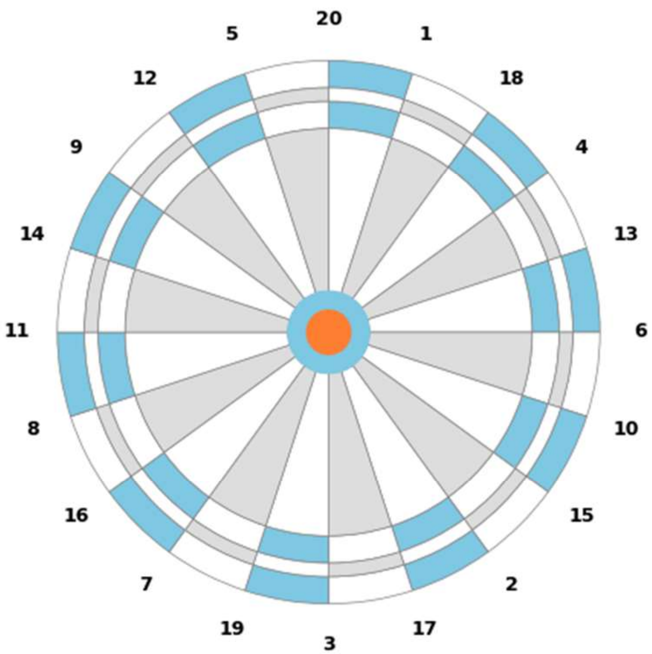


Table 3. Absolute changes in lipid levels, liver values, and albumin values in an analysis that include data on all subjects in the relevant groups.

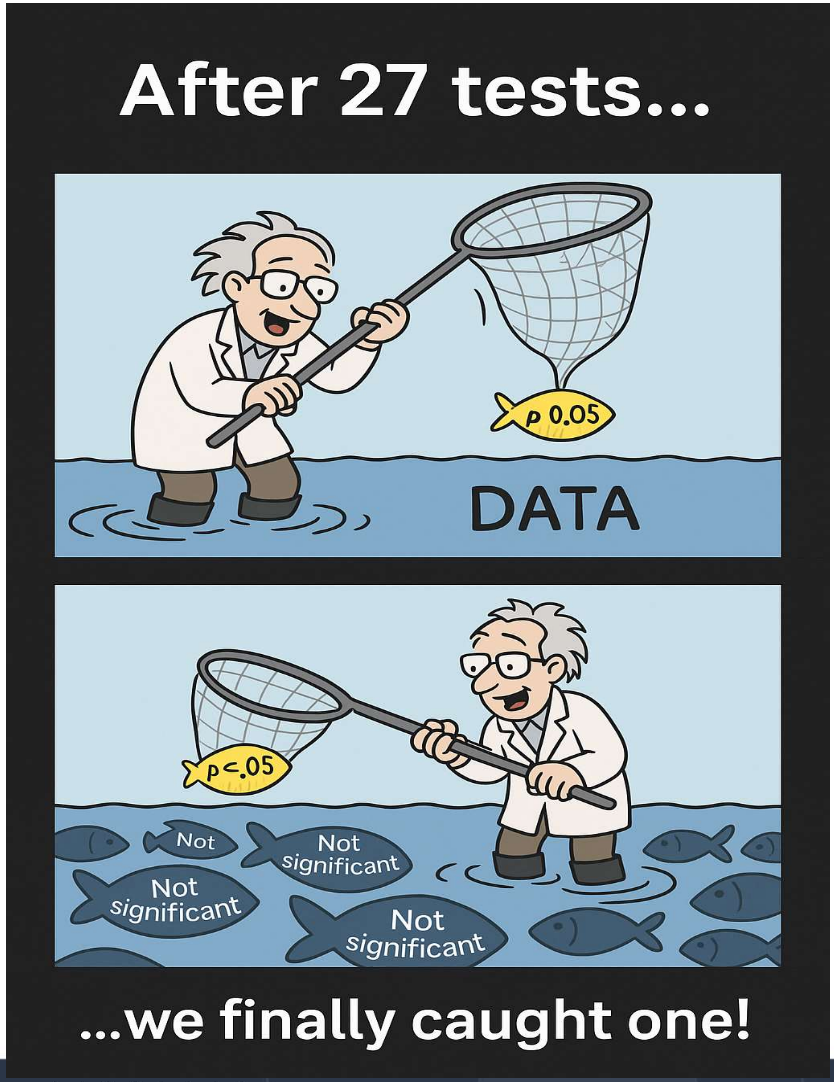
Variable	Chocolate Diet		Low-Carbohydrate		P-Value
Cholesterol (mg/dl). Day 21	-12,2	± 26,7	2,3	± 15,9	0,19
DTriglycerides (mg/dl) Day 21	-22,6	± 85,7	3,0	± 41,3	0,55
LDL cholesterol (mg/dl) Day 21	-17,4	± 22,8	-5,0	± 22,4	0,00
ALT (U/l) Day 21	-6,4	± 6,7	-11,5	± 3,6	0,11
GGT/GGTP (U/l) Day 21	-8,8	± 5,5	-2,0	± 0,0	0.23
Albumin (g/dl) Day 21	0,0	± 0,4	0,1	± 0,3	0.23

Plus-minus values are means ±. The chocolate group had 5 subjects, in the low-carbohydrate group only 4 subjects could be considered.
P values are for the differences between the two groups.

p-hacking

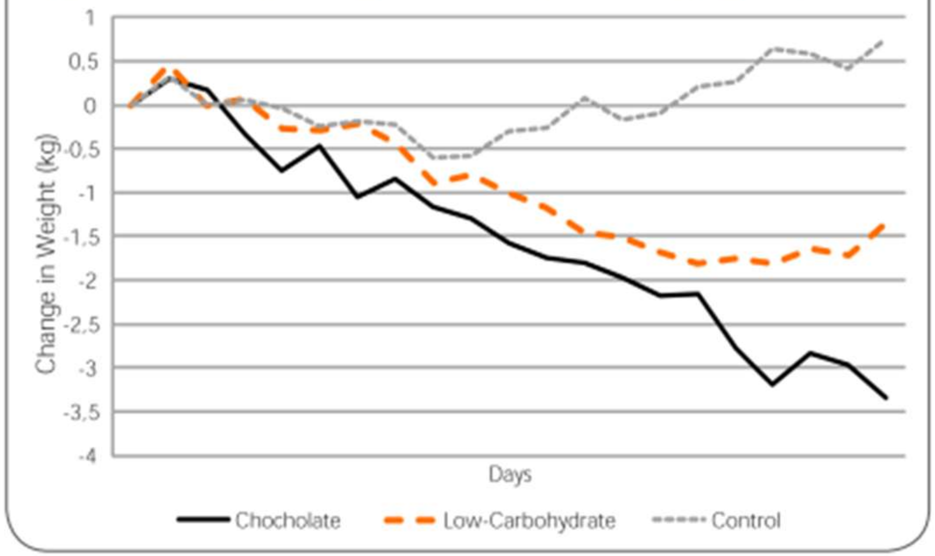
Threshold
α = 5%

Researchers have a lot of degrees of freedom — which subgroups to test, which models to run, which outcomes to highlight. If they keep fishing until something looks ‘significant’, they can almost always find it.



Lack of transparency in the report

Figure 1: Daily weight development by group.



The subjects of the low-carb group lost 3.1 percent of their body weight in 21 days and the chocolate group lost 3.2 percent. The participants of the con-

Replication and generalisability

Coordinated by psychologist Brian Nosek, researchers redid 100 studies in psychological science from three high-ranking psychology journals.
97 of the original studies had significant effects, but of those 97, only 36% of the replications yielded significant findings

- Even if study can be replicated, will it generalise?
- Does it apply beyond the sample studied?
- Will the effect scale in a different context, or at a larger size?

We’ve come a long way in building practices that keep us accountable.

Ways to make ourselves accountable

Pre-register research design and outcomes

Ensures transparency and reduces bias by specifying analysis plans in advance.

Correct for Multiple Testing

Adjusts statistical significance levels to account for multiple comparisons.

Improve transparency in reporting

Don’t hide it in the Appendix

Pre-registration/Pre-analysis plan

AEA RCT Registry

The American Economic Association's registry for randomized controlled trials

About Registration Guidelines Data FAQ Advanced Search

AEA RCT Registry currently lists 10890 studies with locations in 170 countries.

MOST RECENTLY REGISTERED TRIALS

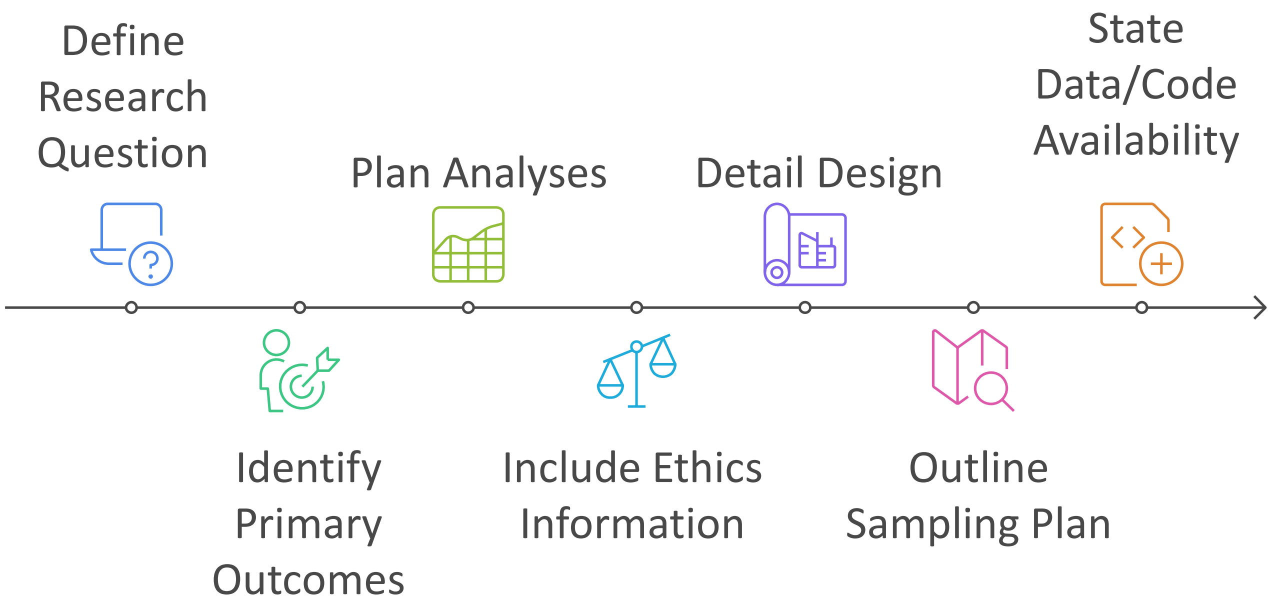
Mindless Entry or Timeless Engagement? An Experimental Study on Strategies to Reduce Screen Time

LAST REGISTERED ON SEPTEMBER 16, 2025

Excessive usage of smartphones poses significant challenges with adverse welfare implications, yet complete restriction is neither feasible nor desirable. This study investigates screen time consumption patterns and the effectiveness of various interventions to reduce excessive smartphone usage. Partnering with one sec, an app that aids in screen time consumption regulation, I conduct a randomized controlled trial (RCT) involving new users to examine pre-entry, post-entry, and combined interventions. The study consists of three phases over four weeks: baseline survey, randomized treatment assignment, and an end-line survey. Key research questions explore the prevalence and interactions of consumption patterns, as well as the intervention adherence and efficacy. By examining the patterns...

Our report has a PAP published as well

What typically goes into pre-analysis plan



CONSORT – guideline on how to present an RCT

CONSORT

TRANSPARENT REPORTING of TRIALS

A checklist

Evaluating claim	•Is this a causal design? (If not, should it claim causality?)
Sample selection	•How were participants selected; who is in the sample? Any selection bias?
Pre-analysis plan	•Was the study pre-registered? Was a primary outcome pre-specified?
Multiple outcomes	•How many outcomes were tested? Were adjustments made for multiple comparisons?
Power	•What is the sample size and the study's statistical power? Could reported results be unstable?

A checklist

Attrition	Were dropouts reported and handled appropriately? Any sensitivity analysis?
Conflict of interest	Is there a conflict of interest or funding source that could influence reporting?
Robustness check	Are alternative explanations (confounding, behaviour change, measurement error) addressed?
Feasible claim	Does the conclusion fairly reflect the evidence, or overstate it?

But there is one final good practice

Program theory and mixed-methods can be incorporated in an impact evaluation

"Why are we seeing the differences we are?

And how did they occur?"

Trial data

Qual data

Additional quant data

etc …

Help to understand the results better, leading to further nuance in the findings

Thank you

Vy.Nguyen@treasury.gov.au