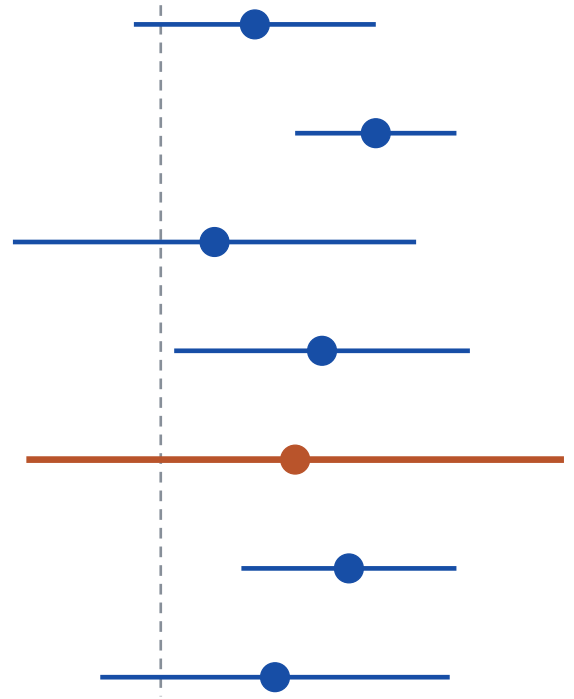
AES26 INTERNATIONAL EVALUATION CONFERENCE · DARWIN

Uncertain impacts, real decisions

Measuring and communicating causal uncertainty in quantitative program evaluations



Dr Dimitria Gavalyugova

Economist, Senior Research and Evaluation Officer at NSW Department of Education

September 2026

NSW Education

ACKNOWLEDGEMENT OF COUNTRY

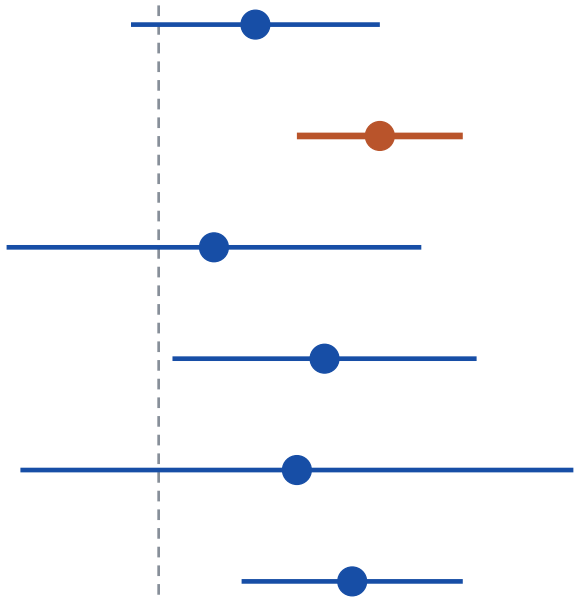
I acknowledge the Larrakia people, the Traditional Custodians of the land on which we meet today, and pay my respects to Elders past and present.

I extend that respect to all Aboriginal and Torres Strait Islander people here today.

01

The evaluator’s dilemma

Decision-makers need clear answers. The evidence often comes with caveats.



Decision-makers expect a definitive impact estimate, but feasible designs rarely deliver one

Strength of evaluation design for causal attribution

Experimental

Units are randomly assigned to the program or not. Groups are comparable on observed and unobserved characteristics, in expectation.

Quasi-experimental

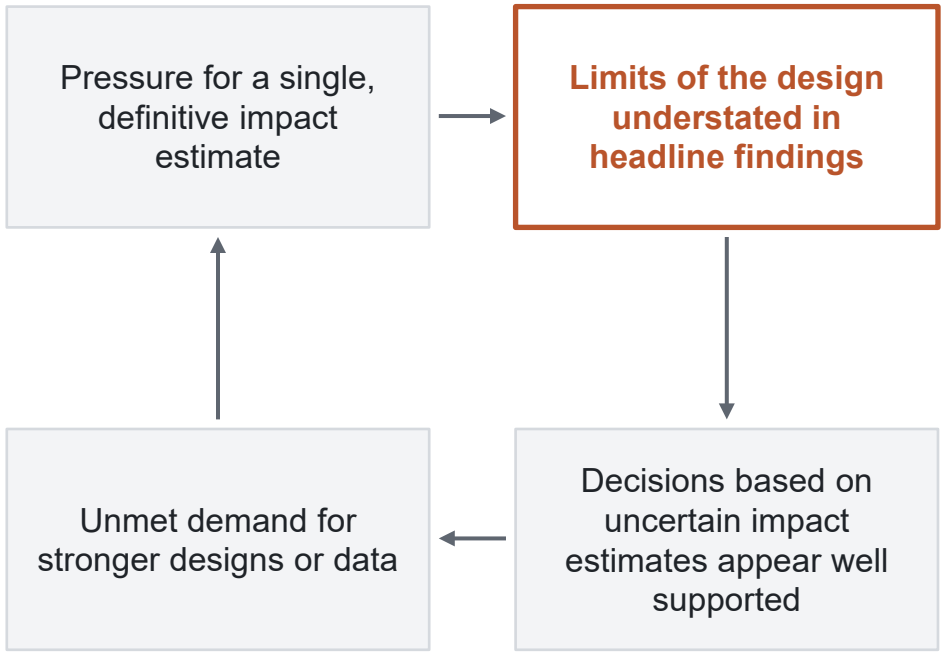
No random assignment. A design feature approximates the counterfactual: e.g. difference-in-differences, regression discontinuity, instrumental variables or matching. Causal claims rest on assumptions such as parallel trends.

Non-experimental

No design to estimate the counterfactual: e.g. before–after or correlational comparisons. Attribution relies on theory, logic and triangulation.

Most large-scale government programs

A sub-optimal equilibrium



Sources: Adapted from NSW Treasury (2023d) and Shadish et al. (2002). Strength depends on how well a design is executed, not only its category. References in appendix.

What I did: mapped causal claims across published NSW evaluations

Sample
Combined outcome and economic evaluations published in the NSW Treasury evaluation library.

Claim strength
For each quantified outcome, found the matching program–outcome claims in the executive summary and in the results section, and coded its language in context.

Evidence strength
Assigned a causal identification level to the design behind each estimate, using a structured strength of causal evidence rubric.

Economic analysis
Checked whether uncertainty in the estimated impact was carried into the cost-benefit analysis.

15
program evaluations

70
outcome estimates scored

54%
of outcomes claimed more strongly in the summary than in the results

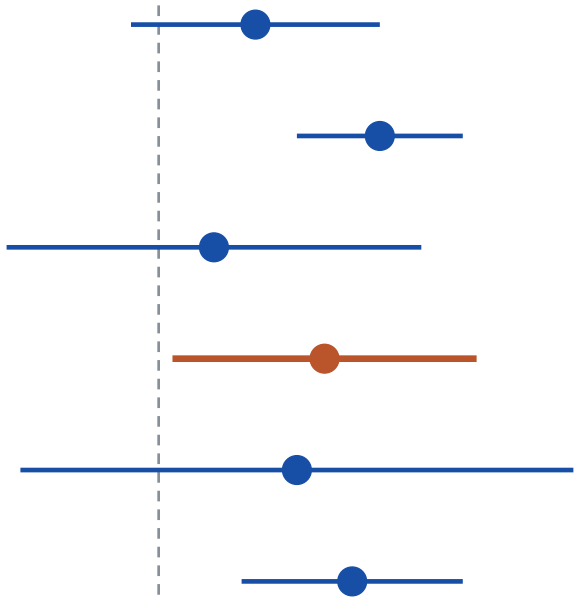
13%
had some uncertainty carried into the CBA

Source: Author-coded structured review (exploratory) of 70 outcome estimates from 15 program evaluations, NSW Treasury evaluation library.

02

Mapping causal claims

How strongly do reports claim a program worked, and how strong is the evidence behind the claim?



I scored each outcome twice: what the report claims, and what the estimate can support

Claim strength · What the report asserts

How strongly the language used attributes the outcome to the program. I coded each program–outcome statement in context in both the executive summary and the results.

Evidence strength · Causal identification

How well the estimation strategy meets the conditions for a causal effect. The hard condition is **internal validity**: ruling out confounding, selection bias and systematic measurement error.

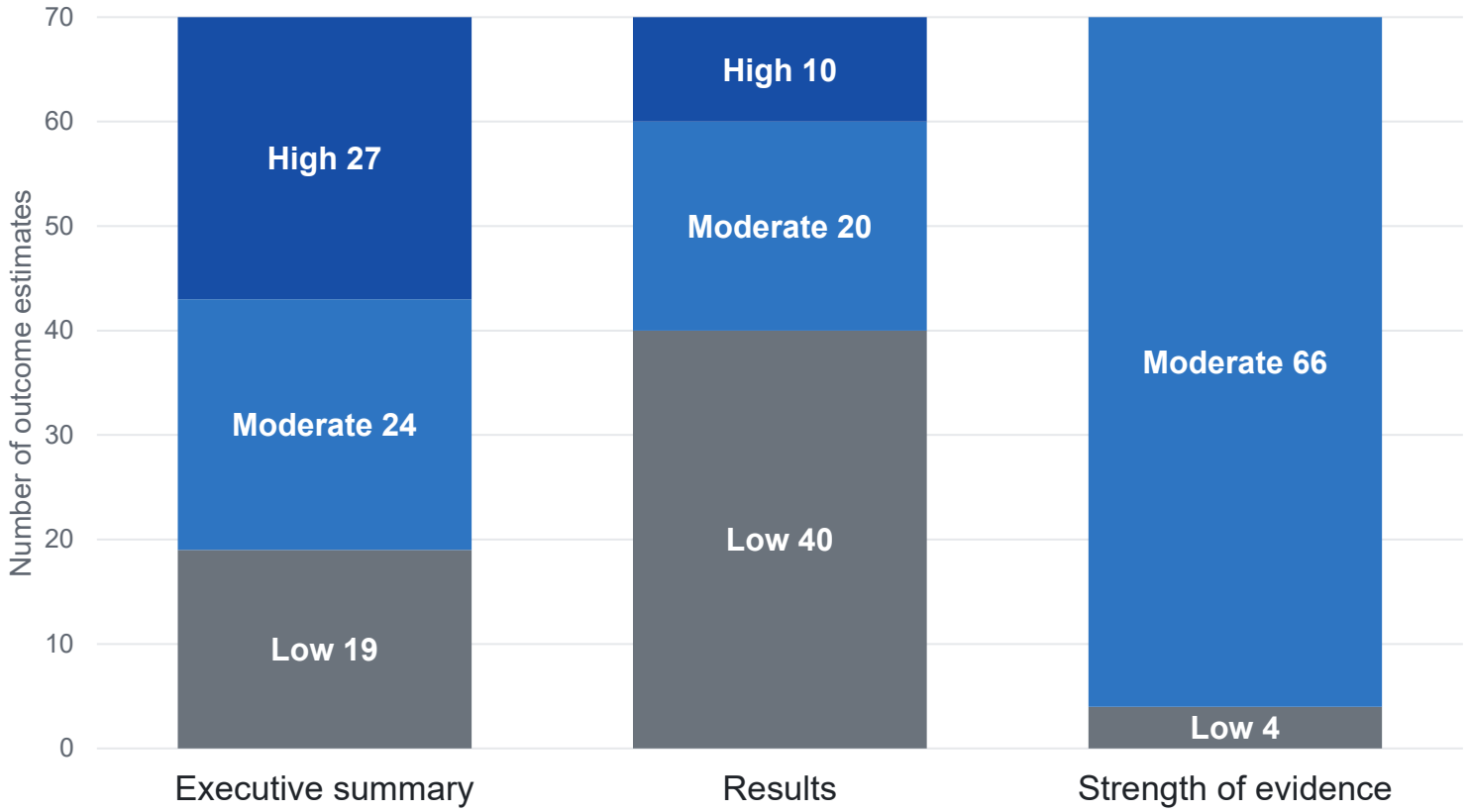
Level	Claim: typical wording	Evidence: conditions met by the estimation strategy
Level 3	Direct causal: “led to”, “improved”, “caused”	Timing, association and internal validity credibly established (e.g. randomised trial, regression discontinuity, strong quasi-experiment with robustness checks)
Level 2	Conditional causal: “contributed to”, “appears to influence”	Timing and association established; internal validity only partly addressed (e.g. quasi-experiment with imperfect comparison group, untestable assumptions)
Level 1	Associational: “associated with”, “linked to”	Timing and statistical association only; no serious attempt to address internal validity

The evidence level is a ceiling. A claim at or below it stays within the evidence. A claim above it exceeds the evidence.

Sources: Language coding adapted from Haber et al. (2022) and Yu et al. (2019). Evidence levels adapted from Campbell (1957), Shadish et al. (2002) and Hernán & Robins (2020).

Executive summaries use stronger causal language than results sections

Claim or evidence level: Low = associational · Moderate = conditional causal · High = direct causal



38 of 70

outcomes are described more strongly in the executive summary than in the results

27

outcomes use direct causal language in the summary

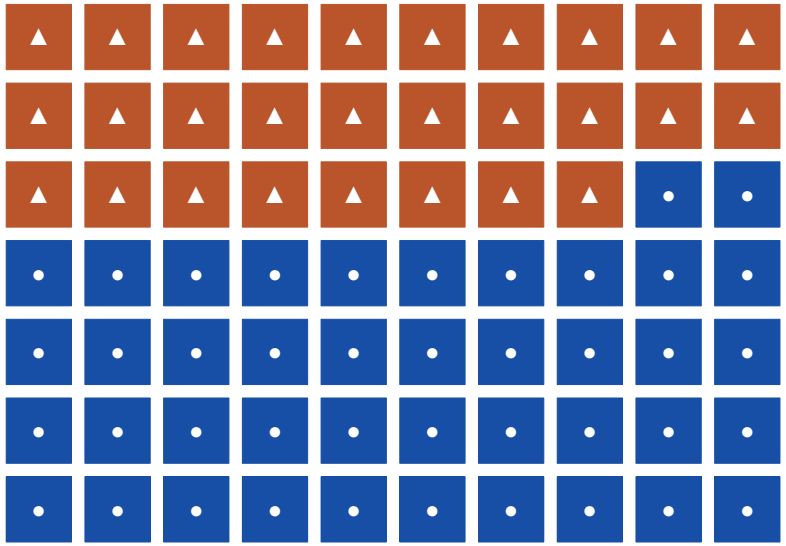
0 of 70

estimates reach the top evidence level (internal validity credibly established)

Source: Author-coded structured review (exploratory) of 70 outcome estimates from 15 program evaluations, NSW Treasury evaluation library.

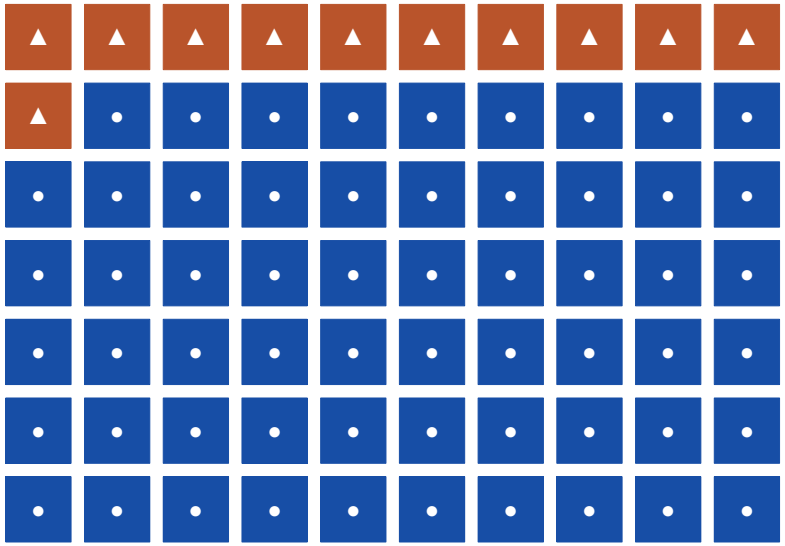
Results sections mostly stay within what the evidence can support; executive summaries often exceed it

Executive summary vs evidence



▲ 28 exceed ● 42 within

Results & methodology vs evidence



▲ 11 exceed ● 59 within

	Summary	Results
▲ Exceeds the evidence	28	11
● Within the evidence	42	59

Each square is one outcome estimate (n = 70).

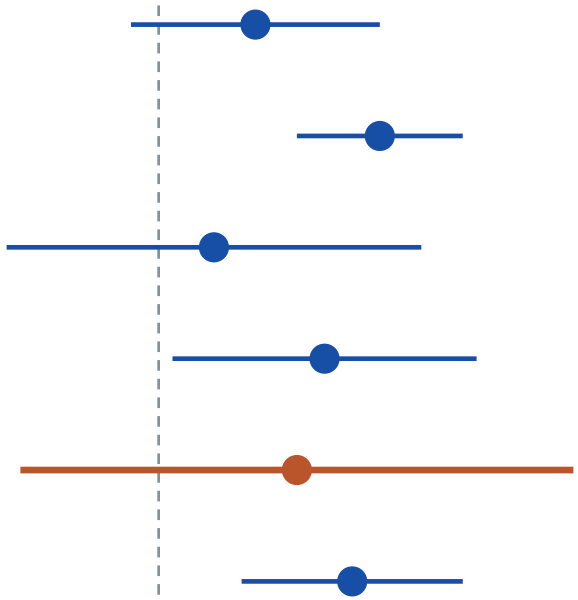
Summaries exceed the evidence for 28 of 70 outcomes (40%). Results sections do so for 11 (16%).

Note: Exceeds = claim level above evidence level. Within = at or below it. Source: Author-coded structured review (exploratory), NSW Treasury evaluation library.

03

From uncertain impacts to benefit-cost ratios

What happens when a potentially biased estimate is monetised, and what we can do about it.



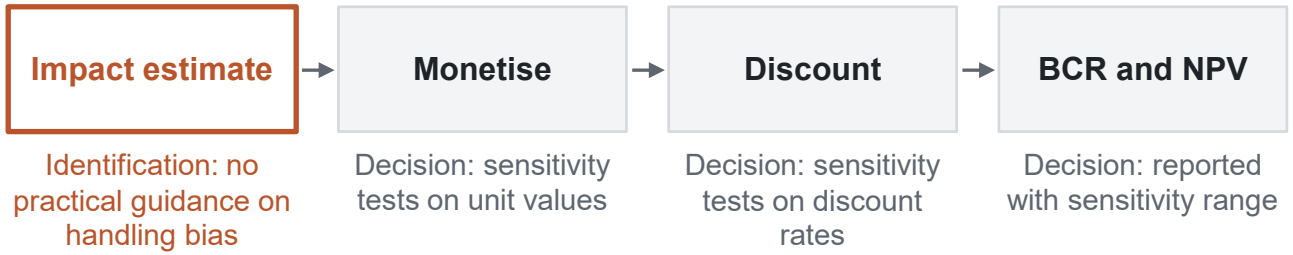
Economic evaluations monetise the whole estimate, including any bias and sampling variability

Estimate = true effect + bias + sampling variability

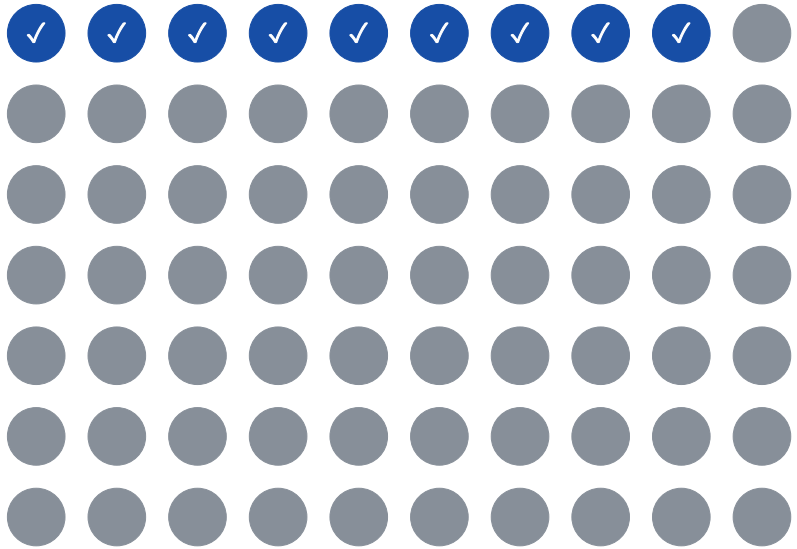


A narrow interval can still be centred on the wrong quantity.

Causal uncertainty: can the design rule out other explanations? Bias from confounding or selection does not shrink with sample size.
Statistical uncertainty: how much would the estimate vary across samples? Shown by standard errors and intervals.



Outcome estimates with statistical uncertainty carried into the CBA



✓ 9 of 70

The other 61 entered as point estimates. Statistical uncertainty is not identification: even the 9 leave possible bias unaddressed.

Sources: NSW Treasury (2023b, 2023c); author-coded structured review (exploratory). References in appendix.

Worked example: from a measured impact to a dollar benefit

The program: Schools offered mentoring and structured activities to help disadvantaged students attend and stay at school. Students chose how involved to get. The evaluation compared students using school records.

1 What they measured

Students who finished Year 12

70% vs 51%

heavily involved vs not involved, in the same schools

A 19-point increase in retention

2 Is all of this gap caused by the program?

↑ Students already likely to finish school may have chosen to get heavily involved.

↑ The report says this comparison cannot establish cause and effect.

3 What the CBA counted

Benefit per participant

$39,100

lifetime earnings, tax and welfare savings

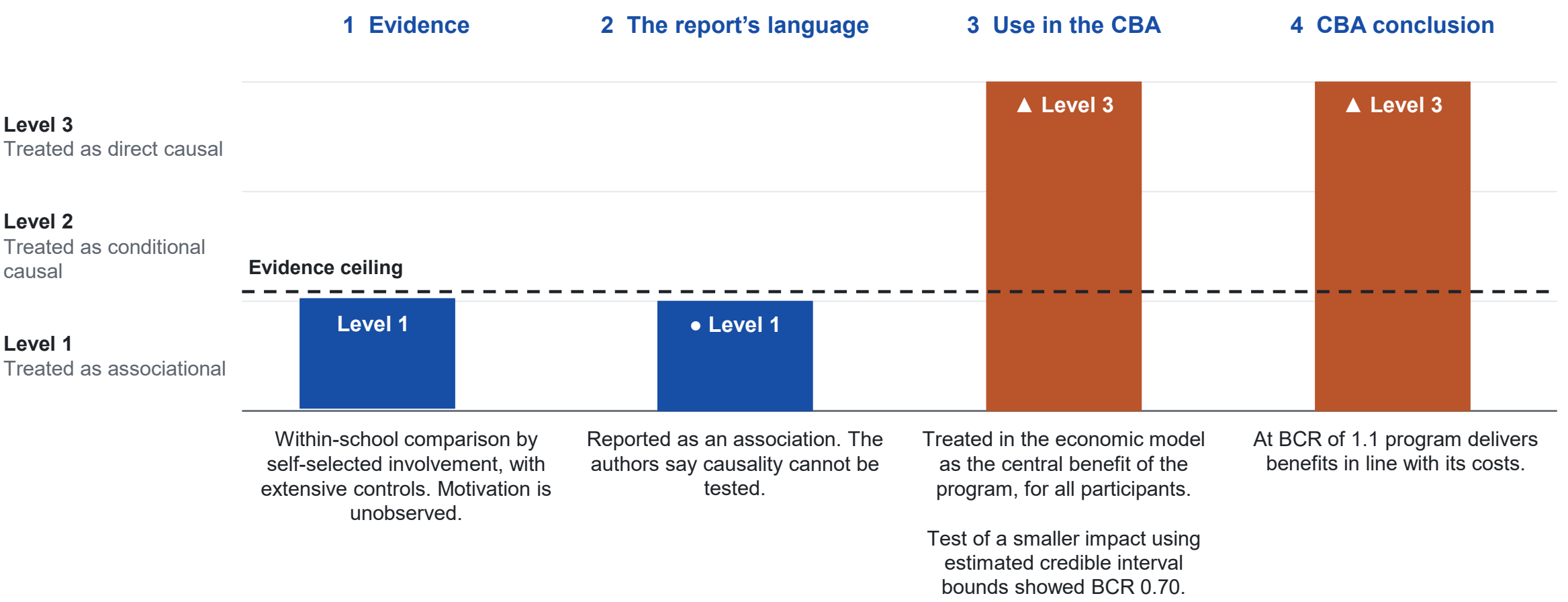
Central BCR 1.10

Source: Anonymised NSW program evaluation. Dollar values adjusted and rounded for anonymity. ↑ = likely to make the gap overstate the program's effect (author's reading of the report).

The gap moved from the words into the numbers

Following the 19-point retention difference through the report

● within the evidence ▲ exceeds the evidence



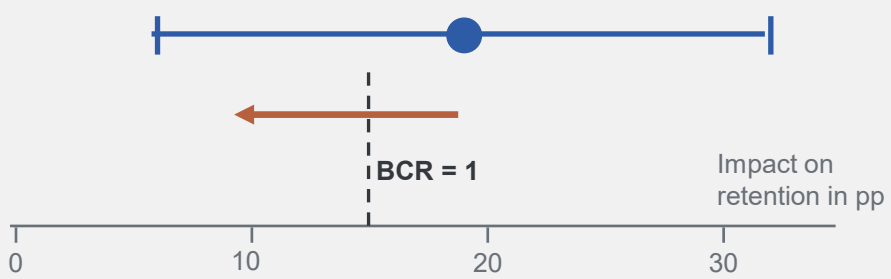
Note: Levels scored by the author with the framework on slide 7; illustrative, not part of the 70-outcome sample. Steps 3 and 4 treat the difference as causal. BCR adjusted for anonymity.

If about a fifth of the measured retention gap was not caused by the program, the BCR falls below 1

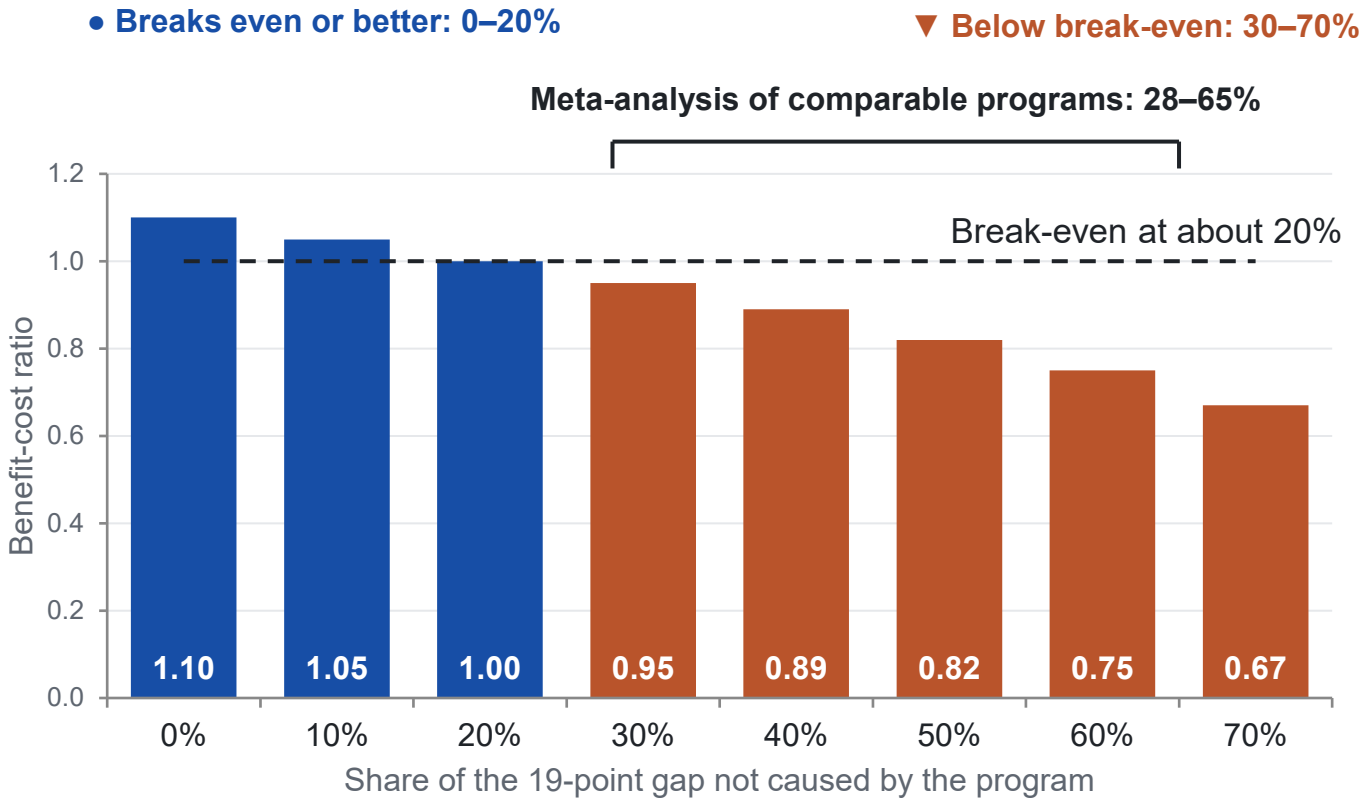
What the bar chart shows

What share of the estimated 19pp change in retention needs to **not be attributable to the program** for the BCR to go below 1?

Why a confidence/credible interval is not enough



Orange arrow: self-selection would shift the whole range. The bounds cannot show this.



Note: Illustrative sensitivity analysis, not an estimate of the true bias. Anonymised, rounded values; 7% real discount rate. Meta-analysis: Hahn et al. (2015); effects of 6.7–13.7 points imply 28–65%. References in appendix.

Four areas to explore together

1 Strengthen the evidence

Which existing data and methods could test the assumptions behind our estimates?

2 Communicate uncertainty

How do we keep headline language matched to what the evidence supports?

3 Model uncertain impacts

Should CBAs report how much bias it would take to change the decision?

4 Design for next time

How can today's limitations become guidance for future evaluation design?

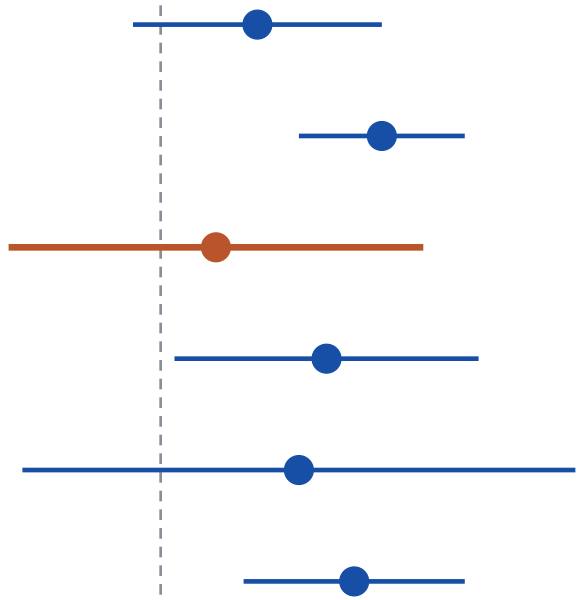
In small groups: think of ideas that fit within those four areas for five minutes. Then share one idea with the room.

Thank you

Let’s connect

Dr. Dimitria Gavalyugova

dimitria.gavalyugova@det.nsw.edu.au



APPENDIX

References

Campbell, D. T. (1957). Factors relevant to the validity of experiments in social settings. *Psychological Bulletin, 54*(4), 297–312. https://doi.org/10.1037/h0040950

Haber, N. A., Wieten, S. E., Rohrer, J. M., et al. (2022). Causal and associational language in observational health research: A systematic evaluation. *American Journal of Epidemiology, 191*(12), 2084–2097. https://doi.org/10.1093/aje/kwac137

Hahn, R. A., Knopf, J. A., Wilson, S. J., et al. (2015). Programs to increase high school completion: A Community Guide systematic health equity review. *American Journal of Preventive Medicine, 48*(5), 599–608. https://doi.org/10.1016/j.amepre.2014.12.005

Hernán, M. A., & Robins, J. M. (2020). *Causal inference: What if*. Chapman & Hall/CRC.

NSW Treasury. (2023a). *Evidence in evaluation* (TPG22-22 Policy and Guidelines: Evaluation – Technical note).

NSW Treasury. (2023b). *Ex-post cost-benefit analysis* (TPG22-22 Policy and Guidelines: Evaluation – Technical note).

NSW Treasury. (2023c). *NSW Government guide to cost-benefit analysis* (TPG23-08; incorporating revisions as at April 2025).

NSW Treasury. (2023d). *Outcome evaluation design* (TPG22-22 Policy and Guidelines: Evaluation – Technical note).

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Houghton Mifflin.

Yu, B., Li, Y., & Wang, J. (2019). Detecting causal language use in science findings. In *Proceedings of EMNLP-IJCNLP 2019* (pp. 4664–4674). Association for Computational Linguistics. https://doi.org/10.18653/v1/D19-1473

Evaluation used in the worked example is not cited, to keep it anonymous.