Choosing and planning ethical evaluations

Session 13

PMAP 8521: Program evaluation
Andrew Young School of Policy Studies

Plan for today

Types of evaluations

Model- and design-based inference

Ethics and open science

Types of evaluations

Types of evaluation

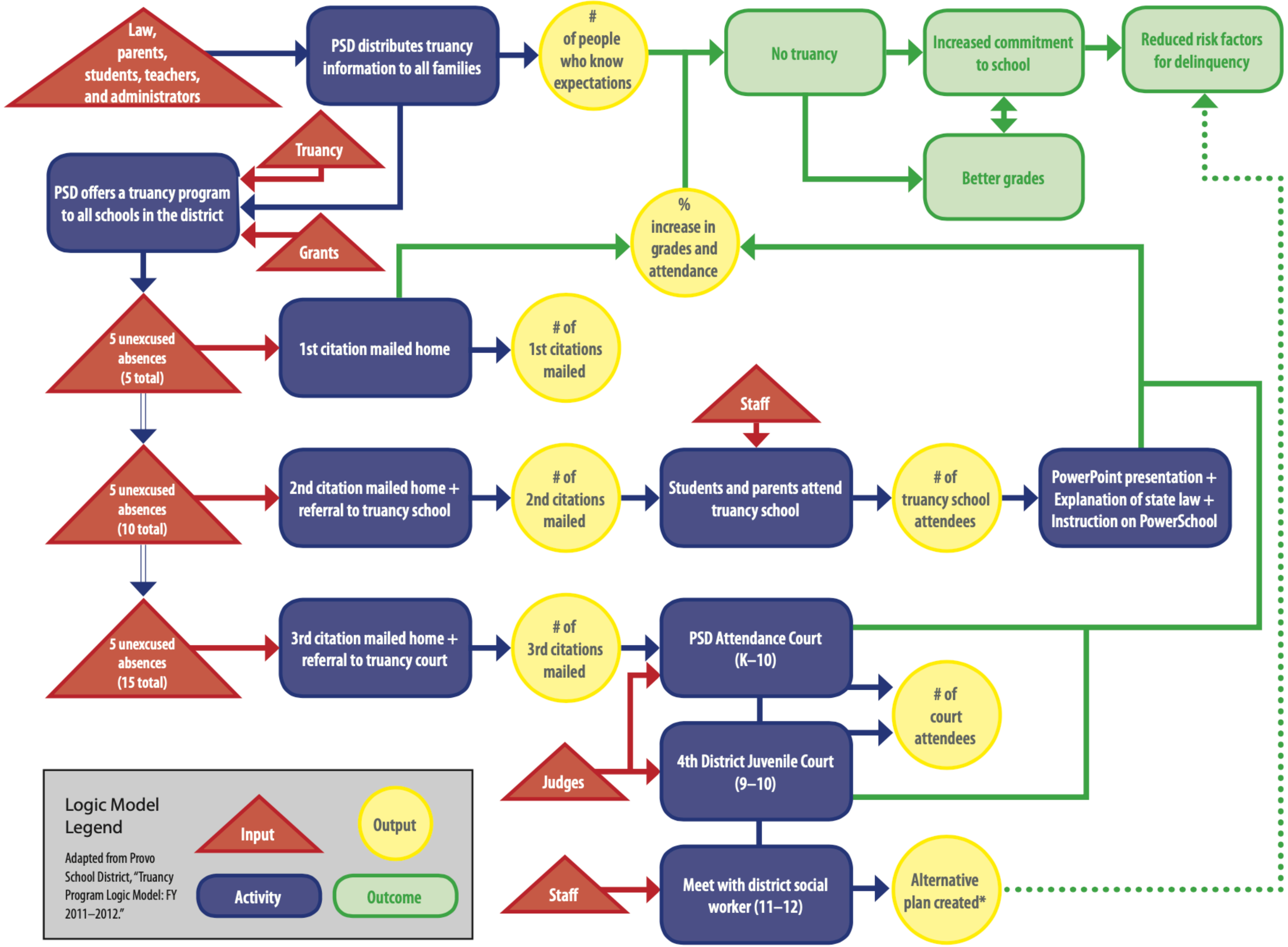
In this class we've focused on one type of evaluation

Impact evaluation

Checking to see if the program causes outcomes

There are lots of others!

Each type focuses on a specific part of a logic model



* Because 11th and 12th graders who receive 3rd citations are generally unable to graduate from high school, district social workers no longer attempt to increase their commitment to school. As such, any outcomes that occur as a result of the alternative plans made for these students (work study programs, career development assistance, etc.) are only tangentially related to the outcomes of the truancy program itself. The system for creating alternative plans is an entirely separate program with its own logic model, goals, and outcomes.

Needs assessment

Formative evaluation / needs assessment

Is the program needed?
What inputs and activities does it need?
What outcomes does it need to cause?

Use interviews, surveys, focus groups with target population

Do *before* starting the program or when considering changes

Process evaluation and monitoring

Process evaluation / program monitoring

Are inputs going to the right places?
Are the activities working correctly?
Are activities producing right levels of outputs?

Use monitoring systems, benchmarks, regular reports from within the program itself

Do *during* the program

Process evaluation and monitoring

Outcome evaluation

Outcome evaluation

Are activities and outputs leading to *initial* outcomes?
(basically a short-term impact evaluation)

Use surveys, interviews, etc. with target population

Do *during* the program

Cost-benefit analysis

Economic evaluation / cost-benefit analysis

Is the program worth it?
Do the benefits of helping the target population outweigh the costs of running the program?

Monetize all program costs and benefits, apply a discount factor, convert all costs to net present value, subtract NPV of costs from NPV of benefits

Do *during* or at the end of the program

Cost-benefit analysis

Table 2
Net Lifetime Benefits of Various Backup Systems
On a Per Vehicle Basis ($2006)

3% discount rate	50 % Driver Factor	80% Driver Factor
Ultrasonic		
At low speeds, 10 % are backing up crashes	-$82.73	-$75.34
At low speeds, 25 % are backing up crashes	-$64.26	-$45.78
Camera		
At low speeds, 10 % are backing up crashes	-$375.21	-$365.20
At low speeds, 25 % are backing up crashes	-$350.19	-$325.16
Both		
At low speeds, 10 % are backing up crashes	-$468.57	-$457.54
At low speeds, 25 % are backing up crashes	-$441.00	-$413.43

7% discount rate	50 % Driver Factor	80% Driver Factor
Ultrasonic		
At low speeds, 10 % are backing up crashes	-$74.23	-$68.35
At low speeds, 25 % are backing up crashes	-$59.53	-$44.83
Camera		
At low speeds, 10 % are backing up crashes	-$365.11	-$357.14
At low speeds, 25 % are backing up crashes	-$345.19	-$325.28
Both		
At low speeds, 10 % backing up	-$447.80	-$439.02
At low speeds, 25 % backing up	-$425.86	-$403.92

Impact evaluation

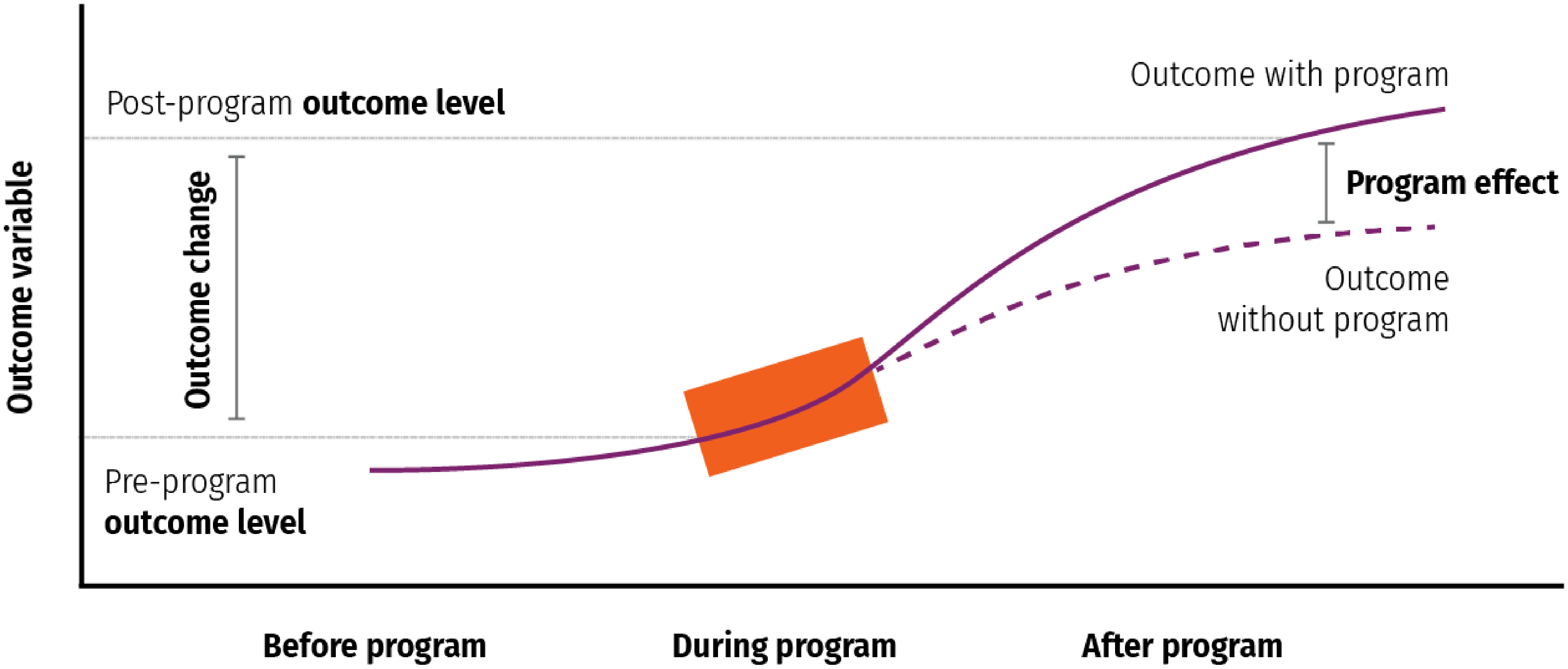
Impact evaluation

Does the program cause lasting change?
(What we did this semester)

Use causal inference tools

Do *during* or at the end of the program

Impact evaluation



Types of evaluation

Needs assessment

Process evaluation and monitoring

Outcome evaluation

Cost-benefit analysis

Impact evaluation

You can take entire classes for just one type!

Model- and design-based inference

Choosing a method

We just learned a *ton* of different methods for causal inference!

DAGs **Matching** **Inverse probability weighting**

Randomized controlled trials **Difference-in-differences**

Regression discontinuity **Instrumental variables**

How do you know which one to use and when?

Identification strategies

The goal of *all* these methods is to isolate
(or **identify**) the arrow between treatment → outcome

Model-based identification

DAGs | **Matching** | **Inverse probability weighting**

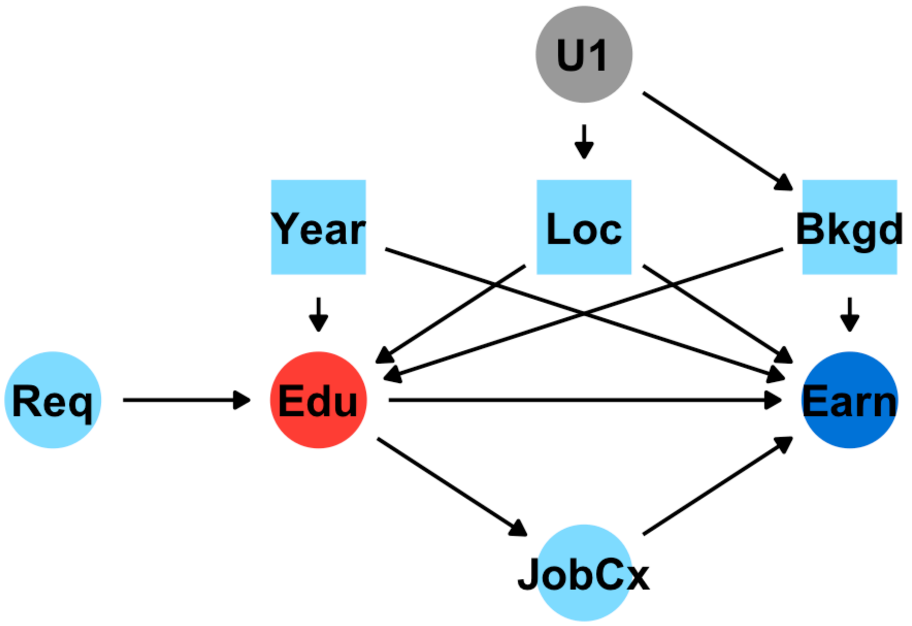
Design-based identification

Randomized controlled trials | **Difference-in-differences**

Regression discontinuity | **Instrumental variables**

Model-based identification

Use a DAG and *do*-calculus to isolate arrow



Core assumption: selection on observables

Everything that needs to be adjusted is measurable; no unobserved confounding

Big assumption!

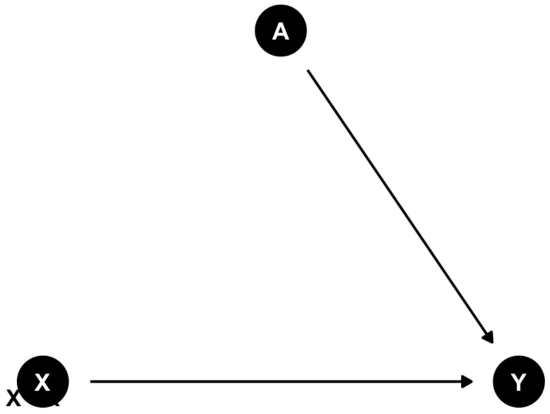
This is why lots of people don't like DAG-based adjustment

Design-based identification

Use a special situation to isolate arrow

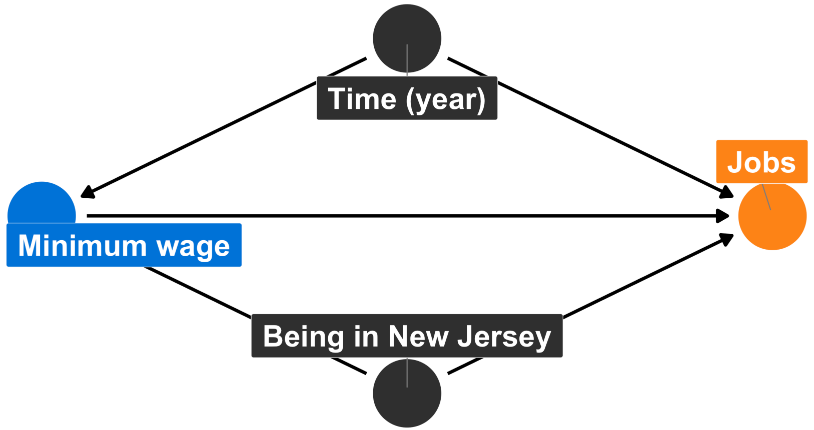
RCTs

Use randomization
to remove confounding



Difference-in-differences

Use before/after & treatment/control
differences to remove confounding

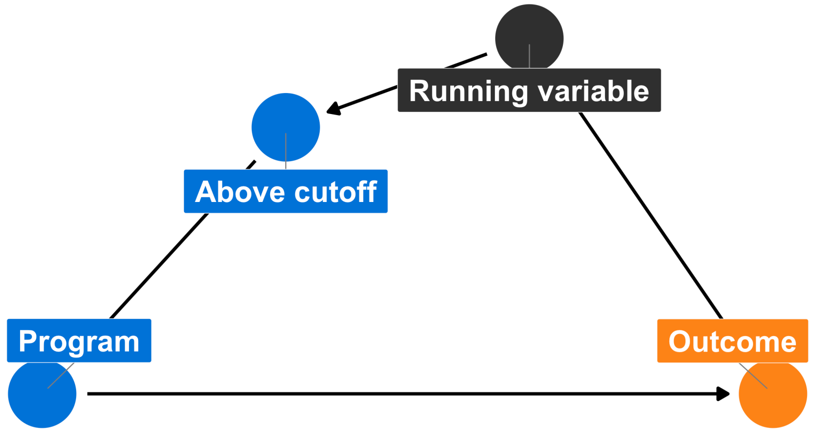


Design-based identification

Use a special situation to isolate arrow

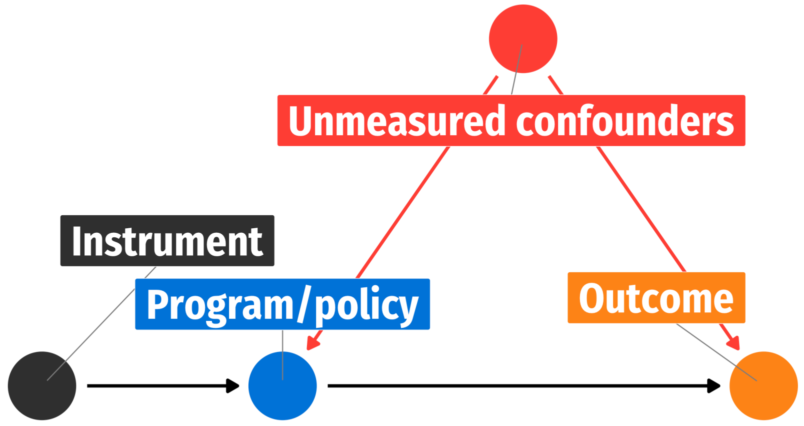
Regression discontinuity

Use cutoff
to remove confounding



Instrumental variables

Use instrument
to remove confounding



Which kind is better?

Model-based advantages

You don't need to wait for
a special circumstance to emerge!

Use existing datasets

Model-based disadvantages

The DAG has to be super correct

You can't adjust your way
out of unobserved confounding

Design-based advantages

Unobserved confounding
is less of a problem!

Design-based disadvantages

You need a specific situation

You need randomization,
treatment/control+before/after,
some arbitrary cutoff,
or some obscure instrument

Controlling for stuff

It's *super* tempting to throw a bunch of control variables in a model

This is likely what you did in past stats classes!

It's *super* tempting to interpret each of those coefficients

Don't!

Table 7: The effect of anti-NGO legislation on the proportion of US aid channeled through *US-based and international* NGOs in the following year (H_3), full models. Each cell contains the parameter's posterior median, the 95% credible interval, and the probability that the parameter is greater than one (in italics)

	(1)	(2)	(3)
Fixed part (odds ratios)			
Total legal barriers$_{within}$	0.95		
	(0.83, 1.08); *0.20*		
Total legal barriers$_{between}$	1.02		
	(0.89, 1.16); *0.60*		
Barriers to advocacy$_{within}$		1.04	
		(0.53, 1.99); *0.54*	
Barriers to advocacy$_{between}$		0.96	
		(0.59, 1.54); *0.44*	
Barriers to entry$_{within}$		1.36	
		(0.98, 1.90); *0.97*	
Barriers to entry$_{between}$		1.07	
		(0.84, 1.35); *0.71*	
Barriers to funding$_{within}$		0.71	
		(0.52, 0.97); *0.01*	
Barriers to funding$_{between}$		0.99	
		(0.76, 1.30); *0.48*	
Civil society reg. env. (CSRE)$_{within}$			1.11
			(0.95, 1.30); *0.89*
Civil society reg. env. (CSRE)$_{between}$			1.03
			(0.89, 1.19); *0.66*
Polity IV (0–10)$_{within}$	1.04	1.04	1.00
	(0.93, 1.18); *0.75*	(0.93, 1.18); *0.74*	(0.87, 1.14); *0.52*
Polity IV (0–10)$_{between}$	0.98	0.98	0.95
	(0.91, 1.06); *0.32*	(0.90, 1.06); *0.30*	(0.84, 1.08); *0.23*
GDP per capita (log)$_{within}$	0.29	0.28	0.28
	(0.17, 0.48); *0.00*	(0.16, 0.47); *0.00*	(0.16, 0.46); *0.00*
GDP per capita (log)$_{between}$	0.72	0.72	0.73
	(0.62, 0.85); *0.00*	(0.62, 0.85); *0.00*	(0.62, 0.85); *0.00*
Trade as % of GDP$_{within}$	1.00	1.00	1.00
	(0.99, 1.00); *0.15*	(0.99, 1.00); *0.14*	(0.99, 1.00); *0.17*
Trade as % of GDP$_{between}$	1.00	1.00	1.00
	(0.99, 1.00); *0.36*	(0.99, 1.00); *0.39*	(0.99, 1.00); *0.36*
Corruption$_{within}$	1.13	1.12	1.16
	(0.96, 1.31); *0.93*	(0.94, 1.31); *0.91*	(0.97, 1.36); *0.95*
Corruption$_{between}$	1.30	1.29	1.30
	(1.19, 1.42); *1.00*	(1.18, 1.42); *1.00*	(1.18, 1.42); *1.00*
Proportion of aid to foreign NGOs in present year (logit)	1.39	1.38	1.39
	(1.33, 1.45); *1.00*	(1.32, 1.45); *1.00*	(1.33, 1.45); *1.00*

Controlling for stuff

When focusing on isolating the treatment → outcome arrow, arrows between/from other nodes are less meaningful

You also don't pick up their full effects!

"[E]ven valid controls are often correlated with other unobserved factors, which renders their marginal effects uninterpretable from a causal inference perspective"

(Hünermund and Louw 2020, p. 2)

Controlling for stuff

Method	Controls	Minimum model
Matching/IPW	Use for matching, propensity scores	`outcome ~ treatment, matched_data` `outcome ~ treatment, weights`
RCTs	Not really necessary	`outcome ~ treatment`
Diff-in-diff	Not really necessary, use if DAG says to	`outcome ~ treatment + after + treatment*after`
RDD	Not really necessary	`outcome ~ running_var + cutoff`
IV	Not really necessary, use if DAG says to	`treatment_hat ~ instrument` `outcome ~ treatment_hat`

Guidelines

Your choice of method depends on the situation + the available data

Table 11.1 Relationship between a Program's Operational Rules and Impact Evaluation Methods

		Excess demand for program (limited resources)		No excess demand for program (fully resourced)	
Timing of Implementation	**Eligibility criteria**	(1) Continuous eligibility ranking and cutoff	(2) No continuous eligibility ranking and cutoff	(3) Continuous eligibility ranking and cutoff	(4) No continuous eligibility ranking and cutoff
	(A) Phased implementation over time	Cell A1 Randomized assignment (chapter 4) RDD (chapter 6)	Cell A2 Randomized assignment (chapter 4) Instrumental variables (randomized promotion) (chapter 5) DD (chapter 7) DD with matching (chapter 8)	Cell A3 Randomized assignment to phases (chapter 4) RDD (chapter 6)	Cell A4 Randomized assignment to phases (chapter 4) Instrumental variables (randomized promotion to early take-up) (chapter 5) DD (chapter 7) DD with matching (chapter 8)
	(B) Immediate implementation	Cell B1 Randomized assignment (chapter 4) RDD (chapter 6)	Cell B2 Randomized assignment (chapter 4) Instrumental variables (randomized promotion) (chapter 5) DD (chapter 7) DD with matching (chapter 8)	Cell B3 RDD (chapter 6)	Cell B4 If less than full take-up: Instrumental variables (randomized promotion) (chapter 5) DD (chapter 7) DD with matching (chapter 8)

Note: DD = difference-in-differences; RDD = regression discontinuity design.

Table 11.1 from *Impact Evaluation in Practice*, p. 191

Ethics and open science

Ethics of evaluating programs

Social programs are designed to help people

In order to evaluate them, you need some people to **not use the program**

Control groups are essential for causal inference!

"Groups should not be excluded from an intervention that is known to be beneficial solely for the purpose of an evaluation"
(*Impact Evaluation in Practice*, p. 233)

Ethical control groups

Table 11.1 Relationship between a Program's Operational Rules and Impact Evaluation Methods

		Excess demand for program (limited resources)		**No excess demand for program (fully resourced)**	
	Eligibility criteria	(1) Continuous eligibility ranking and cutoff	(2) No continuous eligibility ranking and cutoff	(3) Continuous eligibility ranking and cutoff	(4) No continuous eligibility ranking and cutoff
Timing of Implementation	(A) Phased implementation over time	**Cell A1** Randomized assignment (chapter 4) RDD (chapter 6)	**Cell A2** Randomized assignment (chapter 4) Instrumental variables (randomized promotion) (chapter 5) DD (chapter 7) DD with matching (chapter 8)	**Cell A3** Randomized assignment to phases (chapter 4) RDD (chapter 6)	**Cell A4** Randomized assignment to phases (chapter 4) Instrumental variables (randomized promotion to early take-up) (chapter 5) DD (chapter 7) DD with matching (chapter 8)
	(B) Immediate implementation	**Cell B1** Randomized assignment (chapter 4) RDD (chapter 6)	**Cell B2** Randomized assignment (chapter 4) Instrumental variables (randomized promotion) (chapter 5) DD (chapter 7) DD with matching (chapter 8)	**Cell B3** RDD (chapter 6)	**Cell B4** If less than full take-up: Instrumental variables (randomized promotion) (chapter 5) DD (chapter 7) DD with matching (chapter 8)

Note: DD = difference-in-differences; RDD = regression discontinuity design.

Table 11.1 from *Impact Evaluation in Practice*, p. 191

Ethical evaluation practices

Follow IRB guidelines

Respect for persons | **Beneficence** | **Justice**

Make sure participants give informed consent

Maintain privacy

Any published data needs to be be de-identified

Ethical open science practices

Preregistration

Prevents file drawer problem + p-hacking

Replication

Ensures that others can find same results with your data

Preanalysis plan

Prevents p-hacking, data mining, multiple hypothesis testing

Documentation

Ensures that others know what you're measuring

Table 13.1 Ensuring Reliable and Credible Information for Policy through Open Science

Research issue	Policy implications	Prevention and mitigation solutions through open science
Publication bias. Only positive results are published. Evaluations showing limited or no impacts are not widely disseminated.	Policy decisions are based on a distorted body of knowledge. Policy makers have little information on what *doesn't* work and continue to try out/adopt policies that have no impact.	Trial registries
Data mining. Data are sliced and diced until a positive regression result appears, or the hypothesis is retrofitted to the results.	Policy decisions to adopt interventions may be based on unwarranted positive estimates of impacts.	Preanalysis plans
Multiple hypothesis testing, subgroup analysis. Researchers slice and dice the data until they find a positive result for some group. In particular, (1) multiple testing leads to a conclusion that some impacts exist when they do not, or (2) only the impacts that are significant are reported.	Policy decisions to adopt interventions may be based on unwarranted positive estimates of impacts.	Preanalysis plans and specialized statistical adjustment techniques such as index tests, family-wise error rate, and false discovery rate control[a]
Lack of replication. Results cannot be replicated because the research protocol, data, and analysis methods are not sufficiently documented.	Policy may be based on manipulated (positive or negative) results, as results may be due to mistakes in calculations.	Data documentation and registration, including project protocols, organizing codes, publication of codes, and publication of data
Mistakes and manipulations may go undetected.	Results between different studies cannot be compared.	
Researchers are not interested in replicating studies, and journals are not interested in "me-too" results.	Validity of results in another context cannot be tested.	Changes in journal policies and funding policies to require data documentation and encourage replication
Interventions cannot be replicated because the intervention protocol is not sufficiently documented.	Policy makers may be unable to replicate the intervention in a different context.	

a. For a basic introduction to the multiple comparisons problem and potential statistical corrections, please see https://en.wikipedia.org/wiki/Multiple_comparisons_problem.

Table 13.1 from *Impact Evaluation in Practice*, p. 238

Synthetic data

It feels weird to say that making fake data helps with good open science practices!

But it does!

Make your pre-analysis plan based on simulated data

Do whatever statistical shenanigans you want with the fake data