CASSION · COURSE

Impact Evaluation Methods

Counterfactual thinking and the designs that support it — randomisation, difference-in-differences, matching and regression discontinuity — plus how to say what a before-after comparison cannot.

Instructor	Pierre Robentz Cassion
Level	Advanced
Learner hours	24 h
Taught in	Python · R
Updated	July 29, 2026
Sectors	Education · Nutrition · Public Health
Standards and methodologies	OECD DAC evaluation criteria · UNICEF indicator definitions · SMART survey · Theory of Change

Read and run the course online at data-analysis.cassion.dev/courses/impact-evaluation-methods

What you will be able to do

- State the counterfactual a claim requires, and say why a before-after change is not one
- Read a balance table and judge whether a comparison group is one
- Fit a difference-in-differences estimate and argue for the assumption it rests on
- Match on a propensity score and report the units matching discarded
- Recognise an eligibility threshold as a design, and check the three things it needs
- Compute a minimum detectable effect before the evaluation, and say what an underpowered study may conclude

Prerequisites

- Regression for Programme Data

Contents

About this course

Every "report it whole" block in *Regression for Programme Data* ends with the same sentence: *observational, groups were not randomised.* This is the course that says what would have to be true to delete it.

The whole of it can be seen in one question asked three ways of the same file.

Literacy scores rose 7.0 points over the school year. Paired, on 585 children, interval +6.1 to +7.9. A programme document would call that the effect of the school feeding programme.

Schools with a feeding programme end the year 2.1 points ahead of schools without. Interval −1.2 to +5.4. A different document would call *that* the effect.

Both groups gained, and the unfed schools gained more. Fed +6.7, unfed +7.6, so the difference-in-differences estimate is **−0.96 points**, with a school-clustered interval from −3.5 to +1.6. There is no effect here to report, and the first two numbers are the school year and a pre-existing gap wearing an effect's clothes.

Three more results shape the course.

Fifteen schools have the programme and nine do not, and they were not comparable to begin with. The standardised difference in baseline literacy is 0.85, against the 0.1 a randomised trial would tolerate. All six characteristics are imbalanced, and Centre and Nord-Ouest run five fed schools to one unfed.

Matching improves the balance by throwing away 40% of the treatment group. Nine control schools can absorb nine treated schools; the six discarded are the six with the highest baseline scores, and the estimate that survives applies only to the schools that were left.

This design could never have detected anything smaller than 4.9 points. With an intra-cluster correlation of 0.065 and fifty children per school, 24 clusters buy a minimum detectable effect of 4.9 points where the same 1,200 children individually randomised would have bought 2.3. That number was fixed before a single child was measured.

Both Python and R throughout. You need *Regression for Programme Data* first — this course spends no time on how to fit a model and all of it on which comparison the model is allowed to be.

Part I

The counterfactual

CHAPTER 1

A seven-point gain that was the school year

Lesson 1 of 8 · 180 min

Literacy rose 7.0 points between baseline and endline, and a proposal is calling it the effect of the school feeding programme. The schools without the programme rose 7.6 points. All of the gain and slightly more belongs to the year, not the meals.

1.1 One question, three answers

```python
import pandas as pd
import numpy as np

assessment = pd.read_csv("learning-assessment-2024.v1.csv")
assessment["pct"] = assessment["raw_score"] / assessment["items"]
literacy = (assessment[assessment["domain"] == "literacy"]
            .pivot_table(index="student_id", columns="assessment_round",
                         values="pct").dropna())

roster = pd.read_csv("school-roster-2024.v1.csv").drop_duplicates(
    "student_id", keep="last")
d = literacy.join(roster.set_index("student_id"), how="inner")

print(d.groupby("feeding_programme")[["baseline", "endline"]].mean().round(4))
print(f"overall gain: {(d['endline'] - d['baseline']).mean():+.4f}")
```

```r
library(dplyr)

d |> summarise(baseline = mean(baseline), endline = mean(endline),
               .by = feeding_programme)
```

	Baseline	Endline	Gain	n
Feeding programme	57.75%	64.41%	**+6.66 pts**	398
No programme	54.67%	62.29%	**+7.62 pts**	187
All students	56.76%	63.73%	**+6.97 pts**	585

Three claims can be built from that table and two of them are wrong.

"The programme delivered a 7.0-point gain." Before-after, paired, 95% CI +6.1 to +7.9, t = 15.2. Decisive, reproducible, and it attributes to the meals a year of schooling that every child had.

"Schools with the programme are 2.1 points ahead." Endline cross-section, 95% CI −1.2 to +5.4. It compares two groups that were 3.1 points apart before the programme was measured at all.

"The gain was 0.96 points lower in schools with the programme." Both groups against their own baselines, 95% CI −3.5 to +1.6 once the 24 schools are accounted for. **This is the one that is trying to answer the question**, and its answer is that there is nothing here.

1.2 The counterfactual, stated in one sentence

Every impact claim has a hidden second half.

> Literacy in the feeding schools rose 6.66 points **compared with what would have happened without the programme.**

The second clause is the counterfactual, and it is never observed. The same children in the same year without meals is not a thing any dataset contains, so an evaluation is always an argument about what to substitute for it.

Claim	Counterfactual it assumes	Is that plausible?
Before-after	The same children would not have changed at all	**No** — a school year happened
Endline cross-section	The two groups would have been identical	**No** — they differed by 3.1 poi
Difference-in-differences	The two groups would have changed by the same amount	Arguable, and lesson 3 is that

Naming the counterfactual is the whole method. Once it is written down, the before-after claim collapses without any statistics: it assumes children learn nothing in a year they are in school.

1.3 Why the before-after number is so persuasive

It is precise, it is large, and it is easy to compute — three properties that have nothing to do with being right.

```python
gain = d["endline"] - d["baseline"]
se = gain.std(ddof=1) / np.sqrt(len(gain))
print(f"{gain.mean():+.4f}  95% CI [{gain.mean() - 1.96*se:+.4f},"
      f" {gain.mean() + 1.96*se:+.4f}]  t = {gain.mean()/se:.1f}")
```

```r
t.test(d$endline - d$baseline)
```

t = 15.2. Every gate in the statistics course passes. The interval is narrow, the effect size is large, the sample is adequate, and the p-value is far below any threshold.

None of those checks can detect a missing comparison group, and that is the thing worth carrying out of this lesson. Statistical rigour applied to the wrong counterfactual produces a confident wrong answer, and confidence is what gets it into a proposal.

1.4 Three things a before-after change contains

Whenever you see one, it is a sum, and only the last term is the programme.

The passage of time. Children get older and are taught. Prices rise. Water points age. Here it is the entire 7 points.

Regression to the mean. Whoever was selected for being worst off will look better next round even if nothing was done, because part of "worst off" was measurement noise. The nutrition course's admission MUAC is the standing example.

Everything else that happened. A drought, a currency movement, another agency's programme in the same district, a change in how the indicator is recorded.

```python
by_school = d.groupby(["school_id", "feeding_programme"])[
    ["baseline", "endline"]].mean()
by_school["gain"] = by_school["endline"] - by_school["baseline"]
print(by_school.groupby("feeding_programme")["gain"].describe()[
    ["mean", "min", "max"]].round(4))
```

```r
by_school |> summarise(mean = mean(gain), min = min(gain), max = max(gain),
                       .by = feeding_programme)
```

Every one of the 24 schools gained, from +1.2 points to +12.2. A programme that had done nothing at all would have produced a table of positive numbers.

1.5 When a before-after comparison is legitimate

It is not always wrong, and the cases where it holds are worth being precise about.

When the counterfactual is genuinely flat and you can show it. Three or more pre-programme rounds with no trend is an argument; one baseline is not.

When the outcome cannot change on its own. A latrine does not build itself. A water point does not become chlorinated without someone doing it. Coverage of a thing only your programme supplies has a defensible zero counterfactual.

When you are monitoring, not evaluating. "Attendance rose from 84% to 88%" is a true statement about the world and a useful one to track. It becomes a claim about the programme only when someone writes "as a result of".

The sentence to watch for is "as a result of". It is where a monitoring number becomes an impact claim, and it is usually added by someone who was not in the analysis.

1.6 Report it whole

```
Literacy outcomes, baseline to endline 2024

  All students        56.8% to 63.7%    +6.97 points   95% CI +6.1 to +7.9
    With feeding      57.8% to 64.4%    +6.66 points
    Without feeding   54.7% to 62.3%    +7.62 points

  The overall change is reported as a monitoring result and is not
  attributed to the feeding programme. Schools without the programme gained
  slightly more, so the difference-in-differences estimate of the programme
  effect is -0.96 points (95% CI -3.5 to +1.6, clustered on 24 schools).

  The before-after change assumes children would have learned nothing over
  a school year. That counterfactual is not defensible here and the
  comparison group is what replaces it.
```

The last paragraph is three lines and it is the difference between a monitoring report and an impact claim. Writing the counterfactual down is what forces the choice, and a report that never writes one has made the choice by default.

1.7 What comes next

The difference-in-differences estimate above rests on the two groups of schools being comparable in the ways that matter. The next lesson tests that directly, on the fifteen and the nine, and finds a standardised difference eight times what a randomised trial would tolerate.

CHAPTER 2

A standardised difference of 0.85, where 0.1 is the limit

Lesson 2 of 8 · 180 min

Fifteen schools have the feeding programme and nine do not. They differ on baseline literacy by 0.85 standard deviations and on school size by 0.68, and Centre runs five programme schools to one without. This is what a comparison group looks like when nobody randomised.

2.1 Build the balance table before anything else

```python
import pandas as pd
import numpy as np

enrolment = pd.read_csv("school-enrolment-2024.v1.csv")
current = enrolment[(enrolment["school_year"] == 2024)
                    & (enrolment["age_years"] <= 20)]   # drop the age outliers
roster = pd.read_csv("school-roster-2024.v1.csv").drop_duplicates(
    "student_id", keep="last")
joined = current.merge(roster[["student_id", "school_id", "feeding_programme"]],
                       on="student_id", suffixes=("", "_r"))
joined["over_age"] = joined["age_years"] > joined["grade"] + 5
joined["displaced"] = joined["displacement_status"] != "resident"

by = joined.groupby("school_id")
schools = pd.DataFrame({
    "feeding_programme": by["feeding_programme"].first(),
    "baseline": by["baseline"].mean(),      # from the assessment file
    "age_years": by["age_years"].mean(),
    "size": by.size(),
    "over_age": by["over_age"].mean(),
    "disability": by["disability_reported"].mean(),
    "displaced": by["displaced"].mean(),
    "district": by["admin2"].first(),
}).reset_index()

def std_diff(treated, control):
    pooled = np.sqrt((treated.var(ddof=1) + control.var(ddof=1)) / 2)
    return (treated.mean() - control.mean()) / pooled

for column in ["baseline", "age_years", "size", "over_age",
               "disability", "displaced"]:
    t = schools[schools["feeding_programme"]][column].dropna()
    c = schools[~schools["feeding_programme"]][column].dropna()
    print(f"{column:12} {t.mean():8.4f} {c.mean():8.4f}"
          f"  std.diff {std_diff(t, c):+.3f}")
```

```r
library(dplyr)

schools |>
  summarise(across(c(baseline, age_years, size, over_age), mean),
            .by = feeding_programme)
```

Characteristic	Feeding (15)	No programme (9)	Standardised difference
Baseline literacy	54.10%	51.05%	**+0.85**
Mean age	9.65	9.42	**+0.79**
School size	50.3	44.8	**+0.68**
Disability reported	7.6%	9.7%	−0.52
Displaced or returnee	23.7%	26.9%	−0.51
Over-age	39.2%	36.8%	+0.43

The conventional threshold is 0.10, and every one of the six is above it — four of them by a factor of five or more.

```
print(pd.crosstab(schools["district"], schools["feeding_programme"]))
```

```
table(schools$district, schools$feeding_programme)
```

District	No programme	Feeding
Artibonite	3	3
Centre	**1**	**5**
Nord-Ouest	**1**	**5**
Sud	4	2

Two districts are five-to-one in favour of the programme and one is two-to-four against. Whatever else differs between Centre and Sud rides along with every comparison of fed against unfed schools.

2.2 What the standardised difference is, and why not a p-value

```
t = schools[schools["feeding_programme"]]["baseline"]
c = schools[~schools["feeding_programme"]]["baseline"]
from scipy import stats
print(f"std.diff {std_diff(t, c):+.3f}   p = {stats.ttest_ind(t, c).pvalue:.3f}")
```

```
# Both, and only one of them is the right tool.
```

Standardised difference +0.85, p = 0.064. A reviewer reading only the p-value would conclude the schools are balanced on baseline literacy. They are not; the sample is 24 schools and the test has almost no power.

A balance test's p-value confounds imbalance with sample size, which is the opposite of what a balance check needs. Twenty-four units will pass every balance test ever written and two thousand will fail some by chance.

Report the standardised difference, which does not depend on n. Anything above 0.10 is imbalance you have to handle; above 0.25 is imbalance that regression adjustment will not rescue.

2.3 What randomisation would have bought

Not balance on everything — balance *in expectation*, on everything, including the variables you did not measure.

That last clause is the whole value. Adjustment can fix imbalance in baseline literacy because it is in the file. It cannot fix imbalance in how motivated the head teacher is, whether the school was chosen because someone advocated for it, or how far it is from the district office — none of which is recorded anywhere.

```
rng = np.random.default_rng(20260729)
draws = []
for _ in range(2000):
    assigned = rng.permutation(schools["feeding_programme"].values)
    t = schools["baseline"][assigned]
    c = schools["baseline"][~assigned]
    draws.append(std_diff(t, c))
print(f"under random assignment, |std.diff| > 0.85 in "
      f"{np.mean(np.abs(draws) > 0.85):.1%} of draws")
```

```
# Permute the assignment 2,000 times and see how unusual the real split is.
```

Randomly reassigning the same 24 schools produces an imbalance this large in 5.3% of 2,000 draws. So the observed split is unusual but not impossible — which is exactly the wrong conclusion to draw from it. **The question is not whether this imbalance could have arisen by chance; it is that it did not, because nobody randomised.**

2.4 When randomisation is refused, and what to say

It usually is, and the reasons are mostly good ones.

Objection	The honest response
"We cannot deny a service to needy schools"	Randomise the **order** of a phased roll-out; everyone is served, the seq
"The ministry chooses the schools"	Randomise **within** the ministry's eligible list, or evaluate the eligibilit
"It is already running"	Say so, use a comparison design, and report the balance table
"The donor wants results this year"	An underpowered randomised trial and a well-argued difference-in-diff

A phased roll-out is the most under-used design in this sector. Every programme that expands over three years has a randomisable order, and randomising it costs nothing and forecloses none of the objections above.

Where none of that is possible, the balance table is what you owe the reader. It is not a formality: it is the evidence on which they decide how much of the adjustment to believe.

2.5 The three questions to ask before an evaluation design

Who decided who gets the programme, and on what? If the answer is a rule, you may have a threshold design. If it is a person's judgement, you have confounding you cannot enumerate.

Is there anything measured before the programme started? Without a baseline the difference-in-differences of the next lesson is unavailable and you are left with a cross-section.

How many units were assigned? Not how many people were measured — how many schools, clinics, villages, camps. That number sets the precision, and lesson 6 shows what 24 buys.

2.6 Report it whole

```
Comparability of feeding and non-feeding schools

  15 schools with the programme, 9 without. Assignment was not randomised.

  Standardised differences at baseline:
    Baseline literacy     +0.85      School size        +0.68
    Mean age              +0.79      Disability         -0.52
    Displacement          -0.51      Over-age           +0.43

  All six characteristics exceed the 0.10 conventional threshold and all six
  exceed 0.25. Programme schools are concentrated in Centre and
  Nord-Ouest (5 of 6 schools in each) and scarce in Sud (2 of 6).

  Balance tests are not reported as p-values: with 24 units they would show
  balance regardless. The standardised difference does not depend on n.

  The comparison is adjusted for baseline literacy and district. Adjustment
  cannot address unmeasured differences in why these schools were selected,
  and no result in this report should be read as an effect of the programme
  alone.
```

The last paragraph is where an evaluation earns or loses a reviewer, and its length is right: two sentences, stating the limit without apologising for it.

2.7 What comes next

If the two groups differed at baseline, the way to use the baseline is to subtract it. The next lesson does that formally, gets an estimate of -0.96 points, and spends most of its length on the assumption that makes it meaningful — which two rounds of data cannot test.

Part II

When nobody randomised

CHAPTER 3

Minus 0.96 points, and the assumption you cannot test

Lesson 3 of 8 · 180 min

Subtracting each group's own baseline gives a difference-in-differences of −0.96 points. It rests on the two groups having been about to change by the same amount, which is untestable with two rounds — so the lesson is how to argue for it rather than how to check it.

3.1 Two differences, subtracted

```python
import pandas as pd
import numpy as np

cells = d.groupby("feeding_programme")[["baseline", "endline"]].mean()
print(cells.round(4))

did = ((cells.loc[True, "endline"] - cells.loc[True, "baseline"])
       - (cells.loc[False, "endline"] - cells.loc[False, "baseline"]))
print(f"difference-in-differences: {did:+.4f}")
```

```r
library(dplyr)

d |> summarise(baseline = mean(baseline), endline = mean(endline),
               .by = feeding_programme) |>
  mutate(gain = endline - baseline)
```

	Baseline	Endline	Change
Feeding programme	57.75%	64.41%	+6.66
No programme	54.67%	62.29%	+7.62
Difference	+3.09	+2.12	**−0.96**

The estimate can be read down the last column or across the last row and it is the same number. That symmetry is what the name describes: the difference between the two groups' changes, which equals the change in the difference between the two groups.

The baseline gap of 3.09 points is what the design removes. A cross-sectional comparison at endline would have reported +2.12 and called it an effect; the difference-in-differences says most of that gap predates the programme.

3.2 Fit it as a regression, because you need the interval

```python
import statsmodels.formula.api as smf

d = d.assign(gain=d["endline"] - d["baseline"])

naive = smf.ols("gain ~ feeding_programme", data=d).fit()
clustered = naive.get_robustcov_results(cov_type="cluster",
                                        groups=d["school_id"])
```

```python
by_school = d.groupby(["school_id", "feeding_programme"])["gain"].mean().reset_index()
aggregated = smf.ols("gain ~ feeding_programme", data=by_school).fit()
```

```r
lm(gain ~ feeding_programme, data = d)                      # naive
lm(gain ~ feeding_programme, data = by_school)              # school level
```

Approach	Estimate	SE	95% CI
Student level, naive	−0.96 pts	0.98	−2.89 to +0.96
Student level, clustered on school	**−0.96 pts**	**1.29**	**−3.48 to +1.56**
School level, 15 vs 9	−0.68 pts	1.18	−2.99 to +1.63

The clustered interval is the one to report, for the reason the regression course established: the programme was assigned to 24 schools, not to 585 children.

Every version crosses zero comfortably. There is no effect to report, and the interval says the study could not have ruled out anything between a 3.5-point loss and a 1.6-point gain.

3.3 The assumption, stated exactly

Parallel trends: in the absence of the programme, the two groups' outcomes would have changed by the same amount.

Three things that assumption does *not* require, and each is a common misreading:

It does not require the groups to be similar. They start 3.09 points apart and that is fine — the design differences it away. Balance matters for how plausible parallel trends is, not for whether the estimator works.

It does not require the trends to be flat. Both groups can be improving; the assumption is that they would have improved *equally*.

It does not require the same variance, sample size or composition. Those affect the interval, not the identification.

What it does require is unobservable, because it is a statement about a world in which the programme did not happen. That is why this lesson is about argument rather than about testing.

3.4 Why two rounds cannot test it

```python
rounds = assessment["assessment_round"].unique()
print(rounds)
```

```r
unique(assessment$assessment_round)
```

There are two: baseline and endline. With two points you can draw exactly one line through each group, so any pre-programme divergence is unmeasurable by construction.

With three or more pre-programme rounds you can look. Plot each group's outcome over the pre-period and see whether the lines move together. That is not a proof — past parallelism does not guarantee future parallelism — but it is evidence, and it is the single most persuasive exhibit an evaluation of this kind can carry.

The design implication is a data-collection decision made years earlier. If you expect to evaluate by difference-in-differences, collect more than one pre-round. The cost is one extra survey; the alternative is an assumption nobody can examine.

3.5 Four arguments to make when you cannot test it

Each is available here, and together they are what the report offers in place of a test.

Name why the programme schools were chosen. If selection was on something time-invariant — where they are, who runs them — parallel trends is more plausible than if selection was on something trending, like a recent fall in enrolment.

Show that other outcomes moved in parallel. Numeracy is measured on the same children and is not what a feeding programme targets first.

```
def panel(domain):
    wide = (assessment[assessment["domain"] == domain]
            .pivot_table(index="student_id", columns="assessment_round",
                         values="pct").dropna())
    out = wide.join(roster.set_index("student_id"), how="inner").reset_index()
    return out.assign(gain=out["endline"] - out["baseline"])

for domain in ("literacy", "numeracy"):
    frame = panel(domain)
    fit = smf.ols("gain ~ feeding_programme", data=frame).fit(
        cov_type="cluster", cov_kwds={"groups": frame["school_id"]})
    print(f"{domain}: {fit.params['feeding_programme[T.True]']:+.4f}")


# A placebo outcome: it should show nothing, and here it does.
```

Numeracy gives +0.39 points, 95% CI −1.16 to +1.94 — the same design applied to an outcome the programme was not expected to move first, and it shows nothing either. That is weak evidence and it is the kind available.

Test a placebo period if one exists. Two pre-programme rounds should give a difference-in-differences of zero. There are none here, and saying so is better than implying there were.

Show the result is not driven by one unit. Drop each school in turn and refit; if the estimate swings, the design is resting on one school's trend.

```
for school in sorted(d["school_id"].unique()):
    subset = d[d["school_id"] != school]
    est = smf.ols("gain ~ feeding_programme", data=subset).fit()
    print(f"without {school}: {est.params['feeding_programme[T.True]']:+.4f}")


# 24 refits, one line each. Cheap, and a reviewer will ask.
```

Dropping any one school moves the estimate between −1.61 and −0.37 points. No single school carries it, which is worth one line in the annex and is the check a reviewer runs first on 24 units.

3.6 Where difference-in-differences goes wrong

Different timing. If the programme started in different schools in different months, a single before-after split misclassifies part of the treatment period. The staggered case needs a

different estimator, and the naive two-way fixed effects model is now known to be biased for it.

Composition change. The 585 children are those with both rounds. If children left differentially between the groups, the panel is not the same population twice — which is the attrition problem the statistics course's exercise found in this very file.

Anticipation. If schools changed behaviour once they knew the programme was coming, the baseline is already contaminated and the estimate is too small.

A control group that is affected. Children moving between schools, teachers transferring, or a nearby school changing its practice in response all break the assumption that the comparison group shows what would have happened.

3.7 Report it whole

```
School feeding and literacy, difference-in-differences

  585 students with both assessment rounds, in 24 schools.

                       Baseline   Endline   Change
    Feeding (15 sch)     57.75%    64.41%    +6.66
    No programme (9)     54.67%    62.29%    +7.62
    Difference           +3.09     +2.12     -0.96

  Estimate -0.96 points, 95% CI -3.48 to +1.56, clustered on 24 schools.
  No effect on literacy is detected.

  The estimate assumes the two groups would have changed by the same amount
  without the programme. Only two assessment rounds exist, so this cannot be
  examined and is argued rather than tested: assignment appears to follow
  district and school size rather than a trend in results, and the same
  design applied to numeracy also shows nothing.

  The interval excludes effects larger than 1.6 points in either direction
  but not smaller ones. The design's minimum detectable effect was 4.9
  points, so this is a null result about large effects only.
```

The last paragraph converts a null into a statement with a boundary, which is what the statistics course asked for and what the power lesson makes precise.

3.8 What comes next

Difference-in-differences uses the baseline by subtracting it. The next lesson uses it a different way — to choose which comparison schools to keep — and finds that improving the balance means discarding six of the fifteen programme schools.

CHAPTER 4

Better balance, six schools lighter

Lesson 4 of 8 · 180 min

Matching cuts the baseline imbalance from 3.35 points to 1.41 and gives an estimate of −0.71 instead of −0.96. It does it by discarding six of the fifteen programme schools — the six with the highest baseline scores — so the answer now applies to a different set of schools than the question did.

4.1 Matching, in one paragraph

Find, for each treated unit, an untreated unit that looks like it; compare only the pairs. Everything else in the method is a choice about "looks like it" and about what to do when no such unit exists.

The appeal is that it makes the comparison explicit. The cost is that the second half of that sentence — *only the pairs* — changes who the answer is about, and it does so quietly.

4.2 Match the schools

```python
import pandas as pd
import numpy as np

treated = schools[schools["feeding_programme"]].sort_values("baseline")
control = schools[~schools["feeding_programme"]].sort_values("baseline")

used, pairs = set(), []
for _, c in control.iterrows():
    available = treated[~treated["school_id"].isin(used)]
    nearest = (available["baseline"] - c["baseline"]).abs().idxmin()
    used.add(treated.loc[nearest, "school_id"])
    pairs.append({"control": c["school_id"],
                  "treated": treated.loc[nearest, "school_id"],
                  "base_c": c["baseline"], "base_t": treated.loc[nearest, "baseline"],
                  "gain_c": c["gain"], "gain_t": treated.loc[nearest, "gain"]})

matched = pd.DataFrame(pairs)
print(f"{len(matched)} pairs from {len(treated)} treated and {len(control)} control")
```

```r
library(MatchIt)

m <- matchit(feeding_programme ~ baseline, data = schools,
             method = "nearest", ratio = 1)
summary(m)
```

Nine pairs from fifteen treated and nine control schools. One-to-one matching without replacement can produce at most as many pairs as the smaller group has units, so **six treated schools have no match and leave the analysis.**

	Before matching	After matching
Treated schools	15	**9**
Control schools	9	9
Baseline gap	+3.35 pts	**+1.41 pts**
Estimate	−0.96 pts	**−0.71 pts**
Standard error	1.29	1.22

The balance improved and the estimate barely moved, which is the outcome to hope for: it says the difference-in-differences was not being driven by the baseline gap.

4.3 Which six schools left, and why it matters

```
dropped = treated[~treated["school_id"].isin(used)]
print(dropped[["school_id", "baseline", "gain"]].round(4))
print(f"dropped mean baseline: {dropped['baseline'].mean():.4f}")
print(f"kept mean baseline:    {matched['base_t'].mean():.4f}")
```

```
# Which units matching threw away is the first table to print, not the last.
```

The six discarded schools are the six with the highest baseline literacy — 53.9% to 65.5%, against a matched-treated mean of 55.8%. They were dropped because the control group contains no school that scored as highly, so there is nothing to compare them to.

That is not a flaw in the matching; it is the matching telling you something true. The nine control schools cannot speak for the highest-performing programme schools, because no such control school exists.

But it changes the claim. The matched estimate answers "among schools with baseline literacy below about 60%, what did the programme do?" — and that is a narrower question than the one the report set out to answer.

4.4 What "looks like it" should mean

Matching on one variable is a teaching simplification. In practice the choice is between three approaches.

Approach	Match on	Use when
Exact	Identical values of a few categorical variables	Few covariates, larg
Nearest neighbour on a propensity score	The predicted probability of treatment	Many covariates
Coarsened exact	Values binned into ranges	You want a guarant

```
import statsmodels.formula.api as smf

ps = smf.logit("feeding_programme ~ baseline + size + district",
               data=schools).fit(disp=0)
schools["pscore"] = ps.predict(schools)
print(schools.groupby("feeding_programme")["pscore"].describe()[
    ["min", "mean", "max"]].round(3))
```

```
glm(feeding_programme ~ baseline + size + district, data = schools,
    family = binomial())
```

A propensity score is one number summarising many covariates, and matching on it is equivalent to matching on all of them at once — which is what makes it useful when there are more covariates than a small sample can match exactly.

It is not a way to get more data. The score reduces dimensionality; it does not create comparators where none exist.

4.5 Check overlap before you match, not after

This is the check the regression course's exercise found the hard way on the two CMAM programmes, and it applies unchanged here.

```
t = schools[schools["feeding_programme"]]["pscore"]
c = schools[~schools["feeding_programme"]]["pscore"]
print(f"treated range {t.min():.3f}–{t.max():.3f}")
print(f"control range {c.min():.3f}–{c.max():.3f}")
print(f"treated units above the control maximum: {(t > c.max()).sum()}")
```

```
# Plot the two propensity score distributions. The tails are the whole story.
```

Units outside the other group's range have no counterfactual in the data. Matching drops them, which is honest; regression adjustment keeps them and extrapolates, which is not. **That is the real difference between the two methods**, and it is a better reason to prefer matching than any claim about bias.

4.6 Four things to report, always

How many units were discarded, and which. The first table, not a footnote.

Balance before and after. Standardised differences on every covariate, both columns, so a reader can see what matching bought.

What population the estimate now describes. One sentence, in the words a programme manager uses — "schools with baseline literacy below 60%", not "the region of common support".

The unmatched estimate too. If matching changed the answer materially, that is itself a finding about how much the comparison depended on the units it dropped.

4.7 What matching cannot do

It cannot fix an unmeasured confounder. Two schools with identical baseline scores can still differ in why one was selected. Matching balances what you matched on and nothing else — which is exactly the limit of regression adjustment, reached by a different route.

It cannot create a control group. If the untreated units are systematically different, matching will report excellent balance on a handful of pairs and silence about everyone else.

It does not remove the need for a design. Matching plus a baseline is difference-in-differences on a subset; matching alone is a cross-section with better manners. **The estimate above is the first of those**, and combining the two is usually stronger than either.

4.8 Report it whole

```
School feeding and literacy, matched difference-in-differences
```

```
Nearest-neighbour matching on baseline literacy, 1:1 without replacement.
9 matched pairs from 15 treated and 9 control schools.

6 of 15 programme schools were discarded: SCH05, SCH12, SCH14, SCH15,
SCH19, SCH21. Their mean baseline literacy is 60.7% against 55.8% for the
matched treated schools; no control school scored high enough to match
them.

Baseline imbalance   +3.35 points before matching, +1.41 after.
Estimate             -0.71 points, SE 1.22, on 9 pairs.
Unmatched estimate   -0.96 points, 95% CI -3.48 to +1.56.

The matched estimate describes schools with baseline literacy below about
60%. It does not describe the six highest-performing programme schools,
which have no comparator in this data.

Matching balances the covariates it was given. It does not address why
these schools were selected for the programme.
```

The named list of dropped schools is what makes this reportable. A reader can check whether the six that left are the six the programme cares most about, and no summary statistic tells them that.

4.9 What comes next

Matching and difference-in-differences both build a comparison group out of units that were not assigned by a rule. The next lesson takes the opposite case — a programme where a rule decided everything, sharply, at a number — and finds that two of the three things such a design needs are already in the data.

Part III

Rules, and limits

CHAPTER 5

A perfect discontinuity with nothing behind it

Lesson 5 of 8 · 180 min

Every child at 124 mm was referred and 2% of children at 125 mm were. The assignment is as sharp as a rule can be, the density shows no sign of anyone being nudged across, and the design still cannot run — because no outcome was ever measured on the children who were not referred.

5.1 A rule is a design

The MUAC screening protocol refers a child to supplementary feeding below 125 mm and takes no action at or above it. That single number is doing something an evaluation usually has to pay for: **it assigns treatment on a variable, at a known point, with no discretion.**

```python
import pandas as pd
import numpy as np

screening = pd.read_csv("muac-screening-artibonite-2024.v1.csv")
screening["referred"] = screening["outcome"] != "no-action"

window = screening[screening["muac_mm"].between(118, 132)]
print(window.groupby("muac_mm")["referred"].agg(["mean", "size"]).round(3))
```

```r
library(dplyr)

screening |>
  filter(between(muac_mm, 118, 132)) |>
  summarise(referred = mean(outcome != "no-action"), n = n(), .by = muac_mm)
```

MUAC (mm)	Referred	n
122	100.0%	44
123	100.0%	47
124	**100.0%**	**45**
125	**2.0%**	**51**
126	0.0%	78
127	0.0%	70

One millimetre moves the referral probability from 1.00 to 0.02. Children at 124 and 125 mm are, in every way that matters, the same children — the measurement error on a MUAC tape is larger than the gap between them — and one group was treated while the other was not.

That is the closest thing to a randomised experiment that routine data produces, and it arrives free, because the programme was going to apply the rule anyway.

5.2 What a regression discontinuity needs

Three conditions. Two of them are checkable in this file and the third is not.

One: the assignment must actually be sharp at the cut-off. Checked above. If the rule were applied loosely — say 60% referral just below and 20% just above — the design still works but becomes *fuzzy*, and the estimate has to be scaled by the jump in treatment probability rather than read directly.

Two: nobody may manipulate the running variable. If a screener who wants a child treated records 124 instead of 126, the children just below the cut-off are no longer comparable to those just above — they are the ones somebody decided to help.

```python
counts = screening["muac_mm"].value_counts().sort_index()
print(counts.loc[120:130])

below = screening["muac_mm"].between(120, 124).sum()
above = screening["muac_mm"].between(125, 129).sum()
print(f"120-124: {below}   125-129: {above}")

screening["last_digit"] = screening["muac_mm"] % 10
print((screening["last_digit"].value_counts(normalize=True).sort_index() * 100).round(1))
```

```r
table(screening$muac_mm)[as.character(118:132)]
table(screening$muac_mm %% 10)
```

The counts rise smoothly through the cut-off: 45 at 124, 51 at 125, 78 at 126. There is no pile-up just below 125, and the density on the treated side is *lower* than on the untreated side, which is what a genuinely increasing MUAC distribution looks like.

Terminal digits are flat, 9.6% to 10.8% across all ten. No rounding to fives, no avoidance of the cut-off. The data quality course's digit-preference test, used here for a different purpose: **a manipulation check is a digit-preference check asked at one specific number.**

Three: an outcome must be measured on both sides. It is not, and that ends the analysis.

5.3 The condition this register fails

```python
print(screening.columns.tolist())
print(f"children screened once: {(screening['child_id'].value_counts() == 1).sum()}")
print(f"children screened twice: {(screening['child_id'].value_counts() == 2).sum()}")
```

```r
names(screening)
table(table(screening$child_id))
```

The file records a measurement, a referral, and nothing else. 4,194 of 4,206 children appear exactly once. There is no follow-up MUAC, no recovery status, no second screening round — so for the children at 125 mm who were not referred, no outcome exists at all.

The treatment register does not fill the gap. `cmam-admissions-2024` holds outcomes for treated children only, and its identifiers do not link to the screening file. Even if they did, it would carry only the referred side, which is precisely the half a discontinuity design already has.

An impact estimate needs the untreated side. Without it, the sharpest assignment rule in the sector produces nothing.

5.4 What it would have taken

This is the useful part of the lesson, because it is a data-collection specification someone can act on.

A follow-up measurement on children just above the cut-off. Not all of them — children between 125 and 130 mm, re-measured at eight weeks. A few hundred children, one extra visit, and the design becomes estimable.

The same outcome on both sides. Recovery is defined differently for treated and untreated children, and a comparison needs one definition applied to both: MUAC at eight weeks, measured the same way regardless of what happened in between.

Enough children near the cut-off. The estimate uses a window, and everything outside it is discarded.

```
for bandwidth in (2, 3, 5, 10):
    band = screening[
        screening["muac_mm"].between(125 - bandwidth, 124 + bandwidth)]
    print(f"±{bandwidth} mm: {len(band)} children"
          f" ({(band['muac_mm'] < 125).sum()} below, "
          f"{(band['muac_mm'] >= 125).sum()} above)")
```

```
# The bandwidth trade-off, in four lines.
```

Bandwidth	Children	Below cut-off	Above
±2 mm	221	92	129
±3 mm	335	136	199
±5 mm	556	187	369
±10 mm	1,210	265	945

A narrow bandwidth is more credible and less precise. Children within 2 mm of the cut-off are almost identical to each other; there are 221 of them. Widen to 10 mm and you have six times the sample and are comparing children who genuinely differ.

Choose the bandwidth before seeing the outcome, and report the estimate at two or three others. A result that appears only at one bandwidth is a result about the bandwidth.

5.5 Where thresholds hide in programme data

Once you look for eligibility rules, this sector is full of them.

Threshold	Programme	Running variable
MUAC $<$ 125 mm	Supplementary feeding	Arm circumference
MUAC $<$ 115 mm	Therapeutic feeding	Arm circumference
Poverty score below a cut-off	Cash transfer	Proxy means test
IPC Phase 3	Emergency response	Area classification
Below a coverage target	Facility support	Reported coverage

Each is a natural experiment nobody designed, and each needs the same three checks. The third — an outcome on both sides — is the one that most often fails, because programmes measure what they treat.

That is a data-collection decision, not an analysis one. Deciding to measure a sample of ineligible units costs money in a year when nobody is asking for it, and buys an evaluation three years later that would otherwise be impossible.

5.6 Report it whole

```
Supplementary feeding eligibility: assessment of a regression discontinuity

  Assignment: referral to TSFP below MUAC 125 mm.

  Sharpness    100% referred at 124 mm (n=45), 2.0% at 125 mm (n=51).
               The rule is applied without exception.

  Manipulation No pile-up below the cut-off: counts are 45, 51, 78 at 124,
               125 and 126 mm. Terminal digits are uniform (9.5-10.8%).
               No evidence the running variable was adjusted.

  Outcome      Not available. The register records screening and referral
               only; 4,194 of 4,206 children appear once, and no follow-up
               measurement exists on either side of the cut-off.

  Conclusion   The design is available and cannot be estimated. To make it
               estimable, re-measure MUAC at 8 weeks on a sample of children
               screened between 125 and 130 mm, using the same definition
               applied to referred children. 556 children sit within
               +/-5 mm of the cut-off, of whom 369 were not referred.
```

A design assessment that concludes "not estimable" is a deliverable, and this one is more useful than a forced estimate: it names the missing measurement, its size and its cost.

5.7 What comes next

The threshold design failed on data that does not exist. The next lesson is about a design that ran on data that does exist, and asks a question that should have been asked first — how large an effect could it ever have found.

CHAPTER 6

The answer was fixed before the first child was measured

Lesson 6 of 8 · 180 min

Twenty-four schools with fifty children each can detect 4.9 points. The same 1,200 children randomised individually could detect 2.3. The design decided what the evaluation was capable of finding, years before anyone read a result.

6.1 Compute what the design can find

```python
import numpy as np

SD = 0.1429        # within-school SD of attendance, from the register
ICC = 0.065        # intra-cluster correlation, from the empty mixed model
M = 50             # children per school
Z_ALPHA, Z_BETA = 1.96, 0.84       # 5% two-sided, 80% power

deff = 1 + (M - 1) * ICC
print(f"design effect {deff:.2f}")

def mde(k_treated, k_control):
    n1, n0 = k_treated * M, k_control * M
    return (Z_ALPHA + Z_BETA) * SD * np.sqrt(deff * (1 / n1 + 1 / n0))

print(f"15 vs 9 schools: {mde(15, 9) * 100:.2f} points")
```

```
# clusterPower::cpa.normal(), or the formula above. Either way, before the study.
```

Design effect 4.19. Minimum detectable effect 4.88 points.

This evaluation could never have found an effect smaller than about five points on attendance, whatever the analysis did afterwards. That is a property of 24 schools, fifty children each, and an intra-cluster correlation of 0.065 — all three known before any child was measured.

6.2 Where the number comes from

Four inputs, and only two of them are decisions.

Input	Here	Decision or fact?
Outcome standard deviation	0.143	Fact, from a pilot or an earlier round
Intra-cluster correlation	0.065	Fact, and usually the one nobody has
Clusters per arm	15 and 9	**Decision**
Children per cluster	50	**Decision, and the weaker lever**

```python
for k in (12, 24, 40, 60, 100):
    print(f"{k:3} schools: MDE {mde(k // 2, k // 2) * 100:.2f} points")
```

```
# The whole planning conversation in five rows.
```

Schools	Children	MDE
12	600	6.68 pts
24	**1,200**	**4.73 pts**
40	2,000	3.66 pts
60	3,000	**2.99 pts**
100	5,000	2.32 pts

Halving the detectable effect costs roughly four times the schools. Precision improves with the square root, so a design that wants 3 points instead of 4.7 needs 60 schools rather than 24.

6.3 Clusters, not children

This is the number that surprises programme managers and it is the whole reason cluster designs are expensive.

```
individual = (Z_ALPHA + Z_BETA) * SD * np.sqrt(4 / 1200)
print(f"1,200 children, individually randomised: {individual * 100:.2f} points")
print(f"1,200 children in 24 schools:            {mde(12, 12) * 100:.2f} points")
print(f"ratio: {np.sqrt(deff):.2f}")
```

```
# Same children, same outcome, same budget. Twice the detectable effect.
```

2.31 points individually, 4.73 points in clusters — a factor of 2.05, which is the square root of the design effect.

Adding children to existing schools barely helps. With an ICC of 0.065, going from 50 to 100 children per school moves the design effect from 4.19 to 7.44 and the MDE from 4.73 to 4.45 points — a 6% improvement for double the fieldwork.

Adding schools helps in proportion. That is the planning rule: **when the programme is assigned by cluster, buy clusters.**

6.4 The intra-cluster correlation is the input nobody has

It is the hardest of the four to obtain and the one that moves the answer most.

```
for icc in (0.01, 0.03, 0.065, 0.10, 0.20):
    d = 1 + (M - 1) * icc
    m = (Z_ALPHA + Z_BETA) * SD * np.sqrt(d * (1 / 600 + 1 / 600))
    print(f"ICC {icc:.3f}: design effect {d:5.2f}, MDE {m * 100:.2f} points")
```

```
# Vary the ICC across the plausible range and report the MDE for each.
```

ICC	Design effect	MDE, 24 schools
0.01	1.49	2.82 pts
0.03	2.47	3.63 pts
0.065	**4.19**	**4.73 pts**
0.10	5.90	5.61 pts
0.20	10.80	7.59 pts

The MDE ranges from 2.8 to 7.6 points across plausible ICCs. A power calculation quoting one ICC with no range is not a calculation, it is a guess with arithmetic attached.

Where to get one. An earlier round of the same survey, a published study in the same sector and country, or a pilot. Failing all three, report the table above and plan for the pessimistic row — **and say in the protocol which row you planned for.**

6.5 What an underpowered study may conclude

This is the part that changes what gets written, and there are exactly three permitted statements.

"No effect larger than X was detected." With an MDE of 4.9 points and an observed difference-in-differences of −0.96 with an interval from −3.5 to +1.6, the honest sentence is that effects above about 3.5 points in either direction are ruled out and smaller ones are not.

"The study was not designed to detect an effect of the size that matters." If the programme would be considered worthwhile at 2 points and the design detects 4.9, that is a finding about the evaluation, and it should have been found before it ran.

"More units are needed, and here is how many." The table above, with the row a future evaluation should aim at.

Three statements that are not permitted. "The programme had no effect" — the data cannot support it. "The effect was not statistically significant, so the programme should be discontinued" — a decision resting on an absence of evidence. And silence about power, which lets a reader supply the first two for themselves.

6.6 Post-hoc power is not a thing

```
observed = -0.0096
se = 0.0129
print(f"observed effect {observed * 100:+.2f} points, SE {se * 100:.2f}")
```

```
# Do not compute power from the observed effect. It is a rearrangement of p.
```

Power computed from the effect you observed is a monotone function of your p-value — it adds no information and creates a false sense of having checked something. A non-significant result with "low observed power" is exactly a non-significant result.

The useful quantity is the confidence interval, which the statistics course established, and the MDE computed from the *design* rather than from the result. Both are available here; post-hoc power is not one of them.

6.7 Report it whole

```
Statistical power, school feeding evaluation

  Design    15 treated and 9 control schools, ~50 children each.
  Inputs    outcome SD 0.143, ICC 0.065 (estimated from this study's own
            empty mixed model), 5% two-sided, 80% power.

  Design effect 4.19. Minimum detectable effect 4.88 percentage points.
```

```
The observed difference-in-differences is -0.96 points (95% CI -3.5 to
+1.6). Effects larger than about 3.5 points in either direction are ruled
out; smaller effects are not, and the design could not have detected them.

This is reported as a null result about large effects. It is not evidence
that the programme has no effect.

Sensitivity: across ICCs from 0.01 to 0.20 the minimum detectable effect
ranges from 2.8 to 7.6 points. The ICC used is estimated from 24 clusters
and is itself imprecise.

A future evaluation targeting a 3-point effect requires about 60 schools.
Adding children to the existing 24 does not help: doubling to 100 per
school moves the minimum detectable effect from 4.73 to 4.45 points.
```

The last paragraph is the one a programme can act on, and it is the reason to compute power after a null result as well as before a study: it converts "we found nothing" into "here is what finding something would cost".

6.8 What comes next

Every design in this course has been reconstructed after the fact. The next lesson does it in the right order — writing the design, the outcome, the analysis and the threshold of interest down before the data arrives, in a document short enough that someone will actually produce one.

Part IV

What the evaluation claims

CHAPTER 7

One page, written before the data arrives

Lesson 7 of 8 · 180 min

Every design in this course was reconstructed after the fact, and each reconstruction had a choice in it that the result could have influenced. This lesson writes the choices down first — the design, the outcome, the threshold and the analysis — on one page a programme will actually produce.

7.1 Six choices the result could have made for you

Every lesson in this course contained one, and in each case the analyst chose after seeing something.

Lesson	The choice	What it could have been chosen to produce
1	Before-after, cross-section or difference-in-differences	+6.97, +2.12 or −0.96 points
2	Which covariates count as balance	A table that looks acceptable
3	Clustered or naive standard errors	An interval that excludes zero
4	Match or do not match	−0.71 or −0.96, and which schools are in
5	Which bandwidth	Whichever one is significant
6	Which ICC	An MDE below the observed effect

Six binary choices produce sixty-four analyses. None of them is dishonest and any of them can be defended in isolation, which is exactly why the defence has to be lodged before the result is known.

This is the multiple-comparisons problem from the statistics course wearing a different coat. There it was twenty-four tests; here it is one test chosen from sixty-four possible analyses, and the correction is not arithmetic — it is writing the choice down first.

7.2 The page

Not a protocol, not a registered report. **One page, produced in the week the evaluation is commissioned**, because the realistic alternative is nothing at all.

```
Evaluation plan: school feeding and learning outcomes
Written 2024-01-15, before any endline data exists.

1. QUESTION
   Does the school feeding programme raise literacy scores?

2. DESIGN
   Difference-in-differences. 15 programme schools, 9 comparison schools,
   baseline and endline literacy assessment on the same children.
   Assignment was not randomised; the balance table will be reported.

3. OUTCOME
   Percent correct on the literacy assessment. Baseline is out of 40 items
   and endline out of 50, so raw scores are not comparable and percent
   correct is the primary outcome. Numeracy is secondary.
```

```
4. THRESHOLD OF INTEREST
   3 percentage points. Below that the programme would not be expanded on
   these grounds, so an effect smaller than 3 points is reported as
   "no programme-relevant effect" regardless of its p-value.

5. ANALYSIS
   OLS of the individual gain on a programme indicator, standard errors
   clustered on school. Adjusted for baseline score and district.
   Reported with a 95% confidence interval, not a p-value alone.

6. POWER
   ICC assumed 0.065 from the attendance data. Minimum detectable effect
   4.9 points, which is above the 3-point threshold. The evaluation is
   therefore underpowered for the effect size that matters, and this is
   stated in the report whatever the result.

7. WHAT WOULD CHANGE THE CONCLUSION
   Attrition above 20%, or differential attrition between arms above 5
   points, would make the panel unrepresentative and the primary analysis
   would be reported alongside a bounds analysis.

8. WHAT WE WILL NOT CLAIM
   Causality beyond difference-in-differences. Effects on attendance,
   enrolment or nutrition, which are not measured here.
```

Point 6 is the one that makes this document worth writing. It is knowable in January, it says the study cannot answer the question it was commissioned to answer, and finding that out in January costs nothing while finding it out in December costs a year.

7.3 Why each section is there

The question, in one sentence. If it takes more than one, it is more than one evaluation.

The design, named. "We will compare programme and non-programme schools" is not a design; it is three designs that give three answers.

The outcome, defined to the item. The statistics course's exercise found a proposal comparing a 40-item paper to a 50-item paper. Specifying "percent correct" in January is what prevents it.

The threshold of interest, in programme units. This is the section nobody writes and it is the one that stops a significant nothing from being reported as a finding. Ask the programme manager: *how large would this have to be for you to expand it?*

The analysis, specified. Including the standard errors, because the choice between naive and clustered moved the interval in every lesson of the regression course.

The power, computed. Before, not after.

What would change the conclusion. Naming the failure modes in advance means the report does not have to argue about whether they were anticipated.

What you will not claim. The shortest section and the one that protects the evaluation's credibility when someone else over-reads it.

7.4 Deviating from the plan

Plans are wrong, data arrives broken, and the answer is not to pretend otherwise.

Deviations are allowed and must be listed. A short table in the report — what was planned, what was done, why — costs four lines and converts an apparent inconsistency into

a documented decision.

```
Deviations from the evaluation plan

  Planned: adjust for baseline score and district.
  Done:    as planned.

  Planned: primary analysis on all enrolled students.
  Done:    restricted to the 585 students with both assessment rounds.
  Why:     156 students have no endline. Attrition analysis added; the
           students who left scored 39.3% at baseline against 56.8% for
           those who stayed, so the panel is not representative and this
           is reported as a limitation.
```

A listed deviation is a strength. An unlisted one, discovered by a reviewer, is the end of the evaluation's credibility, and the difference between them is four lines written at the time.

7.5 Pre-specification does not mean no exploration

Two sections, clearly labelled. The pre-specified analysis answers the question that was asked. Everything else is exploratory, is labelled exploratory, and generates hypotheses for the next round rather than conclusions for this one.

```
PRIMARY = "gain ~ feeding_programme + baseline + district"   # pre-specified
EXPLORATORY = [                                              # labelled as such
    "gain ~ feeding_programme * sex",
    "gain ~ feeding_programme * baseline_quartile",
    "gain ~ feeding_programme + C(school_id)",
]
print(f"pre-specified: 1 model.  exploratory: {len(EXPLORATORY)} models.")


# Two objects, two headings in the report. That is the whole discipline.
```

Subgroup findings are exploratory unless the subgroup was named in advance. The statistics course found two significant schools out of twenty-four on an effect that was exactly zero; a subgroup analysis chosen after seeing the data is the same machine.

7.6 When there is no plan and the data has arrived

Which is the common case, and it is not hopeless.

Write the plan anyway, dated, before you run the analysis. You have seen the data; you have not yet seen the result. That is worth more than nothing and it is honest to say so.

Pre-specify the primary analysis and report everything else as exploratory. The distinction still means something even when it is drawn late.

State how many analyses you ran. The report block in the statistics course ends with "comparisons run in total", and it belongs here for the same reason.

7.7 Report it whole

```
Analysis plan and deviations
```

```
The evaluation plan was written on 2024-01-15, before endline data
collection, and is reproduced in Annex A.

Primary analysis as pre-specified: difference-in-differences on percent
correct, clustered on school, adjusted for baseline and district.

Deviations: one, listed in Annex A. The panel was restricted to students
with both rounds; an attrition analysis was added.

Exploratory analyses: 3, reported in Annex B and labelled as exploratory.
None is presented as a finding.

Threshold of interest: 3 percentage points, agreed with the programme team
in January. The observed estimate is -0.96 points (95% CI -3.5 to +1.6)
and is below the threshold in magnitude.
```

The date in the first line is the whole document's load-bearing element. Everything else is a claim about intent; the date is what makes it checkable.

7.8 What comes next

The plan says what the evaluation will claim. The last lesson writes the report that results — one whose central finding is that the question it was asked cannot be answered with the data that exists, and which is more useful than the alternative.

CHAPTER 8

The evaluation that declines the question

Lesson 8 of 8 · 180 min

The commission asked whether school feeding raises learning. The honest answer is that this design could not have detected an effect of the size the programme cares about, that what it did detect is nothing, and that both statements have to appear in the summary rather than the annex.

8.1 Assemble what the course produced

Seven lessons, one programme, one file. Here is everything the analysis established.

Question	Answer	Where it came from
Did scores rise?	+6.97 pts, 95% CI +6.1 to +7.9	Before-after, lesson 1
Is that the programme?	No — comparison schools rose 7.62	Lesson 1
Are the groups comparable?	No — six of six characteristics imbalanced	Balance table, lesson 2
What is the effect?	−0.96 pts, 95% CI −3.5 to +1.6	Difference-in-differences, lesson 3
Does matching change it?	−0.71 pts, on 9 of 15 treated schools	Lesson 4
Could a better design run?	Not on this data — no outcome past the cut-off	Lesson 5
What could it have detected?	Nothing below 4.9 points	Power, lesson 6

Two of those seven rows are the report and the other five are the annex. The finding is the fourth row and the seventh, and the seventh is the one that has to lead.

8.2 The summary that is honest

```
School feeding and learning outcomes: evaluation summary

  FINDING

  This evaluation cannot answer whether the school feeding programme
  improves literacy, and the reason is in its design rather than its
  results.

  The programme is delivered in 15 schools and compared against 9. With
  fifty children per school and the observed clustering, the smallest
  effect this comparison could have detected is 4.9 percentage points. The
  programme team's threshold for expansion is 3 points. The evaluation was
  therefore incapable of answering the question it was commissioned to
  answer, and this was knowable before it began.

  What it does establish: literacy rose 7.0 points over the school year in
  both programme and comparison schools, with no difference between them
  (-0.96 points, 95% CI -3.5 to +1.6). Effects larger than 3.5 points in
  either direction are ruled out. Effects between 0 and 3 points -- the
  range the programme cares about -- are not.

  RECOMMENDATION
```

```
Do not discontinue the programme on the basis of this evaluation. Do not
expand it on the basis of the 7-point gain, which is the school year and
is present in schools without the programme.

To answer the question, an evaluation needs about 60 schools rather than
24. If the programme is expanding to new schools, randomising the order of
the roll-out costs nothing and produces a comparison group that does not
have to be argued for.
```

The first paragraph refuses the question and the last paragraph says what would answer it. Between them there is a finding, an interval and a boundary, and no number is presented as an effect that is not one.

8.3 Four sentences an evaluation must be able to write

Each one is a check on the report, and a report that cannot produce all four has a gap.

"The counterfactual is X." Here: schools without the programme, assumed to have been on the same trajectory.

"The estimate is Y, with interval Z, in units the programme uses." Here: −0.96 percentage points of literacy, −3.5 to +1.6, where 3 points is the threshold of interest.

"The smallest effect this design could detect is W." Here: 4.9 points. **This is the sentence most evaluations lack**, and its absence is what lets a null be read as a refutation.

"This analysis cannot rule out V." Here: any effect between zero and about three points, which is the whole range the programme is arguing about.

8.4 What "no effect detected" is worth

It is worth something, and being precise about how much is the difference between a useful null and a damaging one.

Reading	Supported?
"The programme does not work"	**No** — the design could not detect the relevant effect
"The programme has no large effect on literacy"	**Yes** — above 3.5 points is ruled out
"The 7-point gain is not attributable to the programme"	**Yes** — comparison schools gained more
"Funding should be redirected"	**No** — that requires an effect estimate this study does not
"The next evaluation needs 60 schools"	**Yes** — computed, and actionable

The two "yes" rows in the middle are real findings and they are unwelcome ones: they take away a headline number a programme had been using. Delivering that is the job, and doing it alongside the recommendation in the last row is what makes it receivable.

8.5 Writing for the reader who will quote one sentence

Someone will extract one line from this report and put it in a slide. Decide now which line that is.

Not: "No statistically significant effect of school feeding on literacy was found." True, and it will be read as "the programme does not work".

Not: "Literacy rose 7 percentage points in programme schools." True, and it will be read as the effect.

This: "Literacy rose 7 points in both programme and comparison schools; this evaluation was not large enough to detect the 3-point difference the programme cares about."

Put that sentence in the summary, the conclusion and the covering email. The quotable line is chosen by you or it is chosen for you.

8.6 What the course has been about

Eight lessons and one idea: **an impact claim is a comparison with an absent half, and the design is the argument about what fills it.**

Every method here — randomisation, difference-in-differences, matching, discontinuity — is a different argument for the same missing quantity, and each is stronger or weaker depending on facts about how the programme was rolled out rather than on anything in the analysis.

Which means the most consequential decisions were made before any data existed: who got the programme and in what order, what was measured and on whom, and how many units were assigned. An analyst who arrives afterwards is choosing among the arguments the roll-out left available.

So the useful thing this course asks for is not a technique. It is being in the room in January, with one page, asking how the schools will be chosen and whether the order can be randomised — which is a question that costs nothing to ask and is unanswerable a year later.

8.7 Report it whole

```
Evaluation of the school feeding programme: methods and limitations

  Design            Difference-in-differences, 15 programme and 9
                    comparison schools, baseline and endline literacy on
                    585 students with both rounds.

  Counterfactual    Comparison schools, assumed to have been on the same
                    trajectory. Two rounds exist, so parallel trends is
                    argued (numeracy shows the same null; no single school
                    drives the result) and not tested.

  Estimate          -0.96 percentage points, 95% CI -3.5 to +1.6,
                    clustered on 24 schools.

  Power             Minimum detectable effect 4.9 points at 80% power,
                    against a 3-point threshold of interest. Underpowered
                    for the relevant effect size.

  Balance           All six baseline characteristics imbalanced beyond 0.25
                    standardised. Adjusted for baseline and district;
                    unmeasured selection is not addressed.

  Attrition         585 of 741 baseline students have an endline. Those who
                    left scored 39.3% at baseline against 56.8% for those
                    who stayed.

  Not claimed       Any effect on attendance, enrolment or nutrition. Any
                    causal effect beyond the difference-in-differences
                    assumption. Any conclusion about effects below 3 points.

  Plan              Written 2024-01-15, one deviation listed in Annex A,
                    three exploratory analyses in Annex B.
```

Eight rows, and a reviewer can reconstruct every decision from them. That is the deliverable this course has been building toward: not a number, but a number surrounded by everything a reader needs in order to know what it is.

8.8 What comes next

Module 5 is complete: uncertainty, then models, then designs. Module 6 is about delivery — turning what you have established into something a programme can see, use and rebuild, which is the last thing standing between an analysis and a decision.

About this edition

This PDF is generated from the same source as the web version of the course, so the two cannot drift. The web edition carries the runnable code, the downloadable datasets and any corrections issued since this file was produced.

> Every dataset referenced in this course is synthetic or fully de-identified. No record traces to a real person.

This course is published in English and in French. The French edition is a professional adaptation using the terminology UNICEF, WHO, WFP, OCHA and national ministries publish in — not a literal translation.

Read and run the course online at data-analysis.cassion.dev/courses/impact-evaluation-methods

Generated July 29, 2026.