The answer was fixed before the first child was measured

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.

Pierre Robentz Cassion

Lesson 6 of 8

Impact Evaluation Methods — Rules, and limits

What this lesson covers

- Compute what the design can find
- Where the number comes from
- Clusters, not children
- The intra-cluster correlation is the input nobody has
- What an underpowered study may conclude
- Post-hoc power is not a thing
- Report it whole
- What comes next

Compute what the design can find — In 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")
```

Compute what the design can find — In R

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

Compute what the design can find

- 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...

Where the number comes from

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**

Where the number comes from — In Python

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

Where the number comes from — In R

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

Where the number comes from

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

Where the number comes from

- Halving the detectable effect costs roughly four times the schools — Precision improves with the square root, so a...

Clusters, not children — In Python

```
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}")
```

Clusters, not children — In R

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

Clusters, not children

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

The intra-cluster correlation is the input nobody has — In Python

```
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")
```

The intra-cluster correlation is the input nobody has — In R

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

The intra-cluster correlation is the input nobody has

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 intra-cluster correlation is the input nobody has

- The MDE ranges from 2.8 to 7.6 points across plausible ICCs — A power calculation quoting one ICC with no range is not...
- Where to get one — An earlier round of the same survey, a published study in the same sector and country, or a pilot

What an underpowered study may conclude

- "No effect larger than X was detected." — With an MDE of 4.9 points and an observed difference-in-differences of $-0.96\ldots$
- "The study was not designed to detect an effect of the size that matters." — If the programme would be considered...
- "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

Post-hoc power is not a thing — In Python

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

Post-hoc power is not a thing — In R

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

Post-hoc power is not a thing

- Power computed from the effect you observed is a monotone function of your p-value — it adds no information and creates...
- The useful quantity is the confidence interval — which the statistics course established, and the MDE computed from the...

Report it whole — Example (cont.)

```
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
```

Report it whole — Example (cont.)

```
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.
```

Report it whole

- The last paragraph is the one a programme can act on — and it is the reason to compute power after a null result as. . .

What comes next

- Every design in this course has been reconstructed after the fact.

Read the full lesson, with runnable code

https://data-analysis.cassion.dev/courses/impact-evaluation-methods/ie-06-the-effect-you-could-have-detected