A standardised difference of 0.85, where 0.1 is the limit

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.

Pierre Robentz Cassion

What this lesson covers

- Build the balance table before anything else
- What the standardised difference is, and why not a p-value
- What randomisation would have bought
- When randomisation is refused, and what to say
- The three questions to ask before an evaluation design
- Report it whole
- What comes next

Build the balance table before anything else — In Python (cont.)

```
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(),
```

Build the balance table before anything else — In Python (cont.)

```
    "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()
```

Build the balance table before anything else — In Python (cont.)

```
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}")
```

Build the balance table before anything else — In R

```r
library(dplyr)

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

Build the balance table before anything else

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

Build the balance table before anything else

- 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

Build the balance table before anything else — In Python

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

Build the balance table before anything else — In R

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

Build the balance table before anything else

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

Build the balance table before anything else

- Two districts are five-to-one in favour of the programme and one is two-to-four against — Whatever else differs between…

What the standardised difference is, and why not a p-value — In Python

```
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}")
```

What the standardised difference is, and why not a p-value — In R

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

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

- Standardised difference +0.85, p = 0.064 — A reviewer reading only the p-value would conclude the schools are balanced. . .
- A balance test's p-value confounds imbalance with sample size — which is the opposite of what a balance check needs
- Report the standardised difference, which does not depend on n — Anything above 0.10 is imbalance you have to handle;. . .

What randomisation would have bought

- That last clause is the whole value — Adjustment can fix imbalance in baseline literacy because it is in the file

What randomisation would have bought — In Python

```
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")
```

What randomisation would have bought — In R

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

What randomisation would have bought

- Randomly reassigning the same 24 schools produces an imbalance this large in 5.3% of 2,000 draws — So the observed...

When randomisation is refused, and what to say

Objection	The honest response
"We cannot deny a service to needy schools"	Randomise the **order** of a phased roll-out; everyone is served, the sequence is random
"The ministry chooses the schools"	Randomise **within** the ministry's eligible list, or evaluate the eligibility rule as a threshold design
"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-differences are both defensible; a before-after is not

When randomisation is refused, and what to say

- A phased roll-out is the most under-used design in this sector — Every programme that expands over three years has a...
- Where none of that is possible, the balance table is what you owe the reader — It is not a formality: it is the...

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
- Is there anything measured before the programme started? — Without a baseline the difference-in-differences of the next...
- How many units were assigned? — Not how many people were measured — how many schools, clinics, villages, camps

Report it whole — Example (cont.)

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.

Report it whole — Example (cont.)

```
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.
```

Report it whole

- The last paragraph is where an evaluation earns or loses a reviewer — and its length is right: two sentences, stating...

What comes next

- If the two groups differed at baseline, the way to use the baseline is to subtract it.

Read the full lesson, with runnable code

https://data-analysis.cassion.dev/courses/impact-evaluation-methods/ie-02-what-randomisation-buys