Eight points of improvement, two of them nobody earned

Literacy rises from 53.1% to 61.8% of items across the two rounds. On the 585 students who sat both, it rises from 56.8% to 63.7%. The difference is who turned up, and the raw scores could not have told you either number.

Pierre Robentz Cassion

Lesson 7 of 8

What this lesson covers

- The comparison that cannot be made
- Three things that are comparable, and one that is not
- Who sat the test
- Measure the selection instead of assuming it away
- What the panel costs
- Report all three numbers
- What comes next

The comparison that cannot be made — In Python

```
import pandas as pd

assessment = pd.read_csv("learning-assessment-2024.v1.csv")
literacy = assessment[assessment["domain"] == "literacy"]

print(literacy.groupby("assessment_round").agg(
    students=("student_id", "nunique"),
    items=("items", "first"),
    mean_raw=("raw_score", "mean"),
).round(1))
```

The comparison that cannot be made — In R

```r
library(dplyr)

assessment |>
  filter(domain == "literacy") |>
  summarise(students = n_distinct(student_id), items = first(items),
            mean_raw = mean(raw_score), .by = assessment_round)
```

The comparison that cannot be made

Round	Students	Items	Mean raw score
Baseline	741	**40**	21.2
Endline	658	**50**	30.9

The comparison that cannot be made

- A raw score went from 21.2 to 30.9 and that comparison is meaningless — because the instruments are not the same length

Three things that are comparable, and one that is not

- Not comparable: the raw score — Different denominators
- Comparable with care: the proportion of items correct — It puts both rounds on a 0–1 scale, which is necessary and not...

Three things that are comparable, and one that is not — In Python

```
literacy = literacy.assign(proportion=literacy["raw_score"] / literacy["items"])
print(literacy.groupby("assessment_round")["proportion"].mean().round(3))
```

Three things that are comparable, and one that is not — In R

```
assessment |> filter(domain == "literacy") |>
  summarise(proportion = mean(raw_score / items), .by = assessment_round)
```

Three things that are comparable, and one that is not

- 53.1% at baseline and 61.8% at endline — Better, and still resting on an assumption nobody has checked
- Comparable by construction: the proficiency bands — These were equated when the instruments were designed, which is...

Three things that are comparable, and one that is not — In Python

```
bands = pd.crosstab(assessment["assessment_round"],
                    assessment["proficiency_band"], normalize="index")
order = ["below-minimum", "minimum", "proficient", "advanced"]
print((bands[order] * 100).round(1))
```

Three things that are comparable, and one that is not — In R

```
assessment |> count(assessment_round, proficiency_band) |>
  mutate(share = n / sum(n), .by = assessment_round)
```

Three things that are comparable, and one that is not

Band	Baseline	Endline
Below minimum	19.3%	**11.1%**
Minimum	30.6%	23.9%
Proficient	38.2%	38.9%
Advanced	11.9%	**26.1%**

Three things that are comparable, and one that is not

- Below-minimum falls from 19.3% to 11.1% and advanced rises from 11.9% to 26.1% — That is the comparison to report,...

Who sat the test — In Python

```
sat = assessment.groupby("assessment_round")["student_id"].apply(set)
baseline, endline = sat["baseline"], sat["endline"]

print(f"baseline only: {len(baseline - endline)}")
print(f"endline only:  {len(endline - baseline)}")
print(f"both rounds:   {len(baseline & endline)}")
```

Who sat the test — In R

```
# Three groups, and only one of them supports a change estimate.
```

Who sat the test

- 741 sat the baseline, 658 the endline, and 585 sat both — 156 students who sat the first test did not sit the second
- That is not a random 156 — Assessment day catches whoever is in the classroom, and the children who are not in the...

Measure the selection instead of assuming it away — In Python

```
panel = literacy.pivot_table(index="student_id", columns="assessment_round",
                             values="proportion")
panel = panel.dropna()

print(f"panel of {len(panel)} students")
print(f"  baseline {panel['baseline'].mean():.1%} → endline {panel['endline'].mean():.1%}")

all_comers = literacy.groupby("assessment_round")["proportion"].mean()
print(f"all comers")
print(f"  baseline {all_comers['baseline']:.1%} → endline {all_comers['endline']:.1%}")
```

Measure the selection instead of assuming it away — In R

```
assessment |> filter(domain == "literacy") |>
  mutate(proportion = raw_score / items) |>
  select(student_id, assessment_round, proportion) |>
  tidyr::pivot_wider(names_from = assessment_round, values_from = proportion) |>
  filter(!is.na(baseline), !is.na(endline)) |>
  summarise(n = n(), baseline = mean(baseline), endline = mean(endline))
```

Measure the selection instead of assuming it away

	Baseline	Endline	Change
All comers	53.1%	61.8%	**+8.7 points**
Panel of 585	56.8%	63.7%	**+6.9 points**

Measure the selection instead of assuming it away

- Two of the 8.7 points are who turned up — The panel baseline is 3.7 points above the all-comers baseline, which is the...
- Report the panel estimate and show the all-comers figure beside it — The panel is the defensible number; the gap...

What the panel costs

- The panel is not the population — 585 of 741 baseline students, and they are the better-attending ones
- It cannot tell you about the 156 who left — Their learning may have gone anywhere, and the honest report says the. . .

What the panel costs — In Python

```
left = literacy[literacy["student_id"].isin(baseline - endline)
                & (literacy["assessment_round"] == "baseline")]
print(f"the 156 who did not return scored "
      f"{left['proportion'].mean():.1%} at baseline")
print(f"the 585 who did scored "
      f"{panel['baseline'].mean():.1%}")
```

What the panel costs — In R

```
# Quantify the difference rather than asserting it.
```

What the panel costs

- The 156 who did not return scored 39.3% at baseline against the panel's 56.8% — seventeen points below — That is not a...
- Quantifying the selection is the deliverable — not eliminating it

Report all three numbers — Example (cont.)

```
Literacy, baseline to endline

  Proficiency bands, all who sat each round
    Below minimum        19.3% → 11.1%
    Advanced             11.9% → 26.1%

  Proportion of items correct
    All comers           53.1% → 61.8%   +8.7 points   n=741, n=658
    Panel of both rounds 56.8% → 63.7%   +6.9 points   n=585

  Instruments differed: 40 items at baseline, 50 at endline. Raw scores are
  not comparable and are not reported. Proficiency bands were equated at
  design; the proportion of items assumes equivalent difficulty.

  156 baseline students did not sit the endline. They scored 39.3% at
  baseline against 56.8% for those who returned, so the all-comers change
```

Report all three numbers — Example (cont.)

```
overstates learning by about 1.8 points. The panel estimate covers
children who sat both rounds and does not cover those 156.
```

What comes next

- You now have every education indicator this course can produce — enrolment, attendance, retention and learning — each with a denominator and a caveat.

Read the full lesson, with runnable code

https://data-analysis.cassion.dev/courses/education-programme-analysis/edu-07-comparing-two-rounds