Five dimensions, five measures

Calling the data poor is not a finding. Accuracy, completeness, timeliness, consistency and integrity — each with a number attached, computed on the same extract, and a scorecard that survives being disagreed with.

Pierre Robentz Cassion

Lesson 1 of 8

What this lesson covers

- The sentence to stop saying
- The five, and the measure for each
- Compute all five on one extract
- Weight the dimensions by what they cost
- Do not collapse it to one number
- What comes next

The sentence to stop saying

- "The data quality is poor." Everyone nods, nothing happens, and the same sentence is said next quarter.

The five, and the measure for each

Dimension	The question	The measure
Completeness	Did everyone report, on everything?	Reporting rate; missing-value rate per field
Timeliness	Did the report arrive in time to be used?	Share submitted by the deadline; median days late
Accuracy	Does the reported figure match the source?	Verification factor: recount over reported
Consistency	Do the numbers agree with each other and with last month?	Rate of failed internal rules; outlier rate
Integrity	Was the value produced by measurement or by something else?	Digit preference; heaping; implausibly stable series

The five, and the measure for each

- Accuracy is the only one that needs the source document — The other four are computable from the extract you already. . .
- Integrity is not an accusation — It measures whether numbers look like measurements

Compute all five on one extract — In Python (cont.)

```
import pandas as pd

vax = pd.read_csv("vaccination-coverage-2024.v1.csv", parse_dates=["period"])
vax["reported"] = vax["report_submitted"] == True

scorecard = {}

# Completeness
scorecard["reporting_rate"] = vax["reported"].mean()
scorecard["field_completeness"] = 1 - vax.isna().mean().mean()

# Consistency: doses cannot exceed the target population
rule_fail = (vax["doses_administered"] > vax["target_population"]).mean()
scorecard["rule_failure_rate"] = rule_fail

# Integrity: share of reported values ending in zero
```

Compute all five on one extract — In Python (cont.)

```
reported = vax[vax["reported"]]
scorecard["round_number_share"] = (reported["doses_administered"] % 10 == 0).mean()

print(pd.Series(scorecard).round(3))
```

Compute all five on one extract — In R

```r
library(dplyr)
library(readr)

vax <- read_csv("vaccination-coverage-2024.v1.csv")

scorecard <- tibble::tibble(
  reporting_rate     = mean(vax$report_submitted),
  field_completeness = 1 - mean(is.na(as.matrix(vax))),
  rule_failure_rate  = mean(vax$doses_administered > vax$target_population),
  round_number_share = mean(
    vax$doses_administered[vax$report_submitted] %% 10 == 0
  )
)

round(scorecard, 3)
```

Compute all five on one extract

Measure	Value	Read as
Reporting rate	76.5%	642 of 2,736 facility-months carry no report
Field completeness	100%	no blank cells — which is not the same as no missing data
Rule failure rate	0%	no facility reports more doses than its target population
Round-number share	10.4%	what chance produces; nothing to see

Weight the dimensions by what they cost

- For a **coverage figure**, completeness dominates. A 76.5% reporting rate makes any district total a lower bound, and...
- For a **caseload figure used for procurement**, accuracy dominates. Ordering therapeutic food against an over-reported...
- For an **early warning indicator**, timeliness dominates. A perfectly accurate report arriving six weeks after the...
- Say which dimension you weighted and why, in the report — A composite quality score that hides the weighting is the...

Do not collapse it to one number — In Python

```
scorecard_table = pd.DataFrame({
    "dimension": ["Completeness", "Timeliness", "Accuracy", "Consistency", "Integrity"],
    "measure": ["Reporting rate", "Submitted by deadline", "Verification factor",
                "Rule failure rate", "Round-number share"],
    "value": [0.765, None, None, 0.0, 0.104],
    "source": ["extract", "submission log", "facility visit", "extract", "extract"],
})
print(scorecard_table)
```

Do not collapse it to one number — In R

```
scorecard_table <- tibble::tribble(
  ~dimension,     ~measure,                ~value, ~source,
  "Completeness", "Reporting rate",         0.765, "extract",
  "Timeliness",   "Submitted by deadline",     NA, "submission log",
  "Accuracy",     "Verification factor",       NA, "facility visit",
  "Consistency",  "Rule failure rate",       0.000, "extract",
  "Integrity",    "Round-number share",      0.104, "extract"
)
```

Do not collapse it to one number

- The `source` column is the useful one — It says immediately which numbers you can produce today and which need someone...

What comes next

- Two of the five dimensions are almost never reported at all, and both are computable from data you already hold.

Read the full lesson, with runnable code

https://data-analysis.cassion.dev/courses/data-quality-assessment/dqa-01-five-dimensions