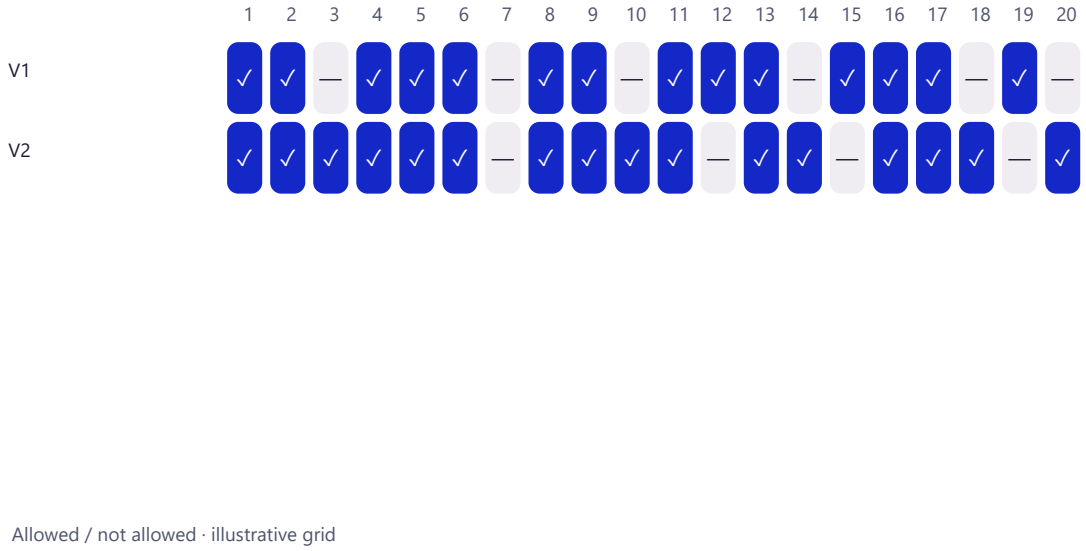
CHAPTER 19 / EVALUATION

One eval set, different graders.

See how the grader and the presentation order change the conclusion.

See how the grader and the presentation order change the conclusion.



Allowed / not allowed · illustrative grid

Synthetic fixture. Repeating cases changes the arithmetic illustration but supplies no new evidence.

Rubric

Define observable criteria and representative cases before scoring.

Dimensions

Report accuracy, grounding, format, usefulness and safety separately.

Regression

Compare the same cases; investigate weak slices rather than hiding them in one average.

↗ **Define observable criteria and representative cases before scoring.**

PUT IT TO WORK

Grade the same 20 synthetic cases by hand, by exact match and with a scripted judge in both answer orders.

CLOSE THE PAGE. EXPLAIN IT.

Which conclusion survives changing the grader or reversing the answer order?