Science 9–10

Chapter 5 — Evaluating

Aboriginal and Torres Strait Islander content has been excluded as accuracy cannot be guaranteed, and it is better to be respectful than cover the entire curriculum inaccurately.

Significant effort has been made to ensure this content is legal. Please send any areas of concern to admin@vwx.au with appropriate document/legal references so errors can be verified and changes be made.

Contents

Validity, reproducibility, random and systematic error, and areas of uncertainty

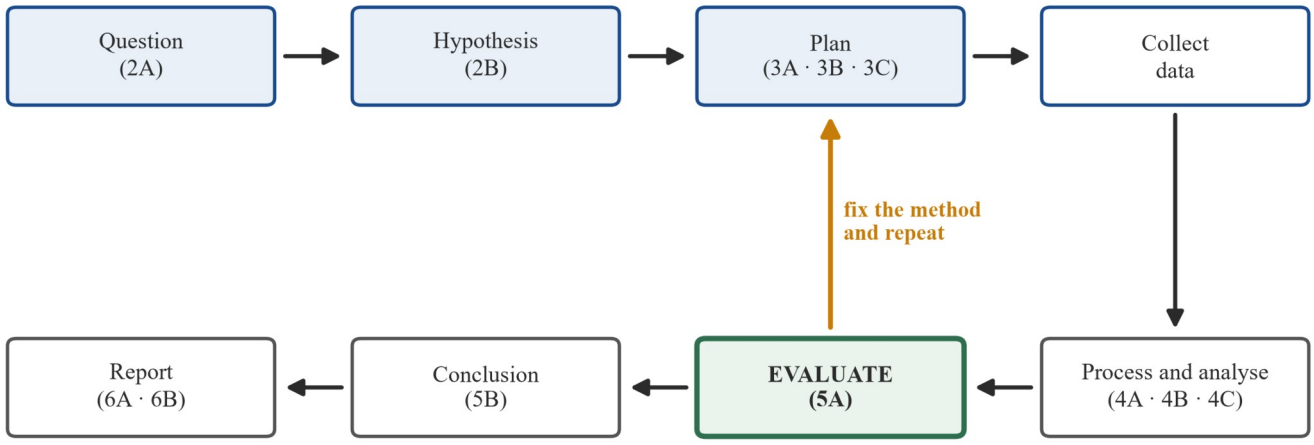
Chapter 5 — Evaluating (Science, Levels 9 and 10)

> **Subtopic 5A** · VC2 Science Inquiry Skills — Evaluating
>
> **CD code:** `VC2S10I06` (delivered whole by this subtopic) **Content description (verbatim):** "the validity and reproducibility of investigation methods and the validity of conclusions and claims can be evaluated, including by identifying assumptions, conflicting evidence, biases that may influence observations and conclusions, sources of error and areas of uncertainty" **Elaboration ids drawn on:** `VC2S10I06.E01` · `VC2S10I06.E02` · `VC2S10I06.E03` **Achievement standard (verbatim):** "They evaluate the validity and reproducibility of investigation methods including ways to improve the quality of data, and the validity of conclusions and claims." **Maths interlocks (D11):** `VC2M9M04` — calculate and interpret absolute, relative and percentage errors in measurements (the size of a systematic shift can be stated as an error); `VC2M9ST01` / `VC2M9ST02` — how samples are chosen and how data is obtained affect survey results and the point of view a display supports (the sampling and bias checks here use that idea).

All datasets in this subtopic are simulated data for practice.

Theory

Evaluation — the step that judges the whole investigation. Every investigation in this book follows the same inquiry sequence: a question, a hypothesis, a plan, collected data, processing and analysis — and then **evaluation**, the step this chapter is about. Evaluation looks back over everything that came before it and asks: *was the method valid? could someone else reproduce it? how strong is the conclusion the data actually supports?* It is not a summary of the results; it is a judgement about how much trust the evidence deserves. And it is not the end of the sequence either — what evaluation finds feeds straight back into the plan, because the point of judging a method is to fix it and repeat.



The inquiry sequence of this book. Evaluation judges the method, the data and the claim — and its findings feed straight back into a better plan.

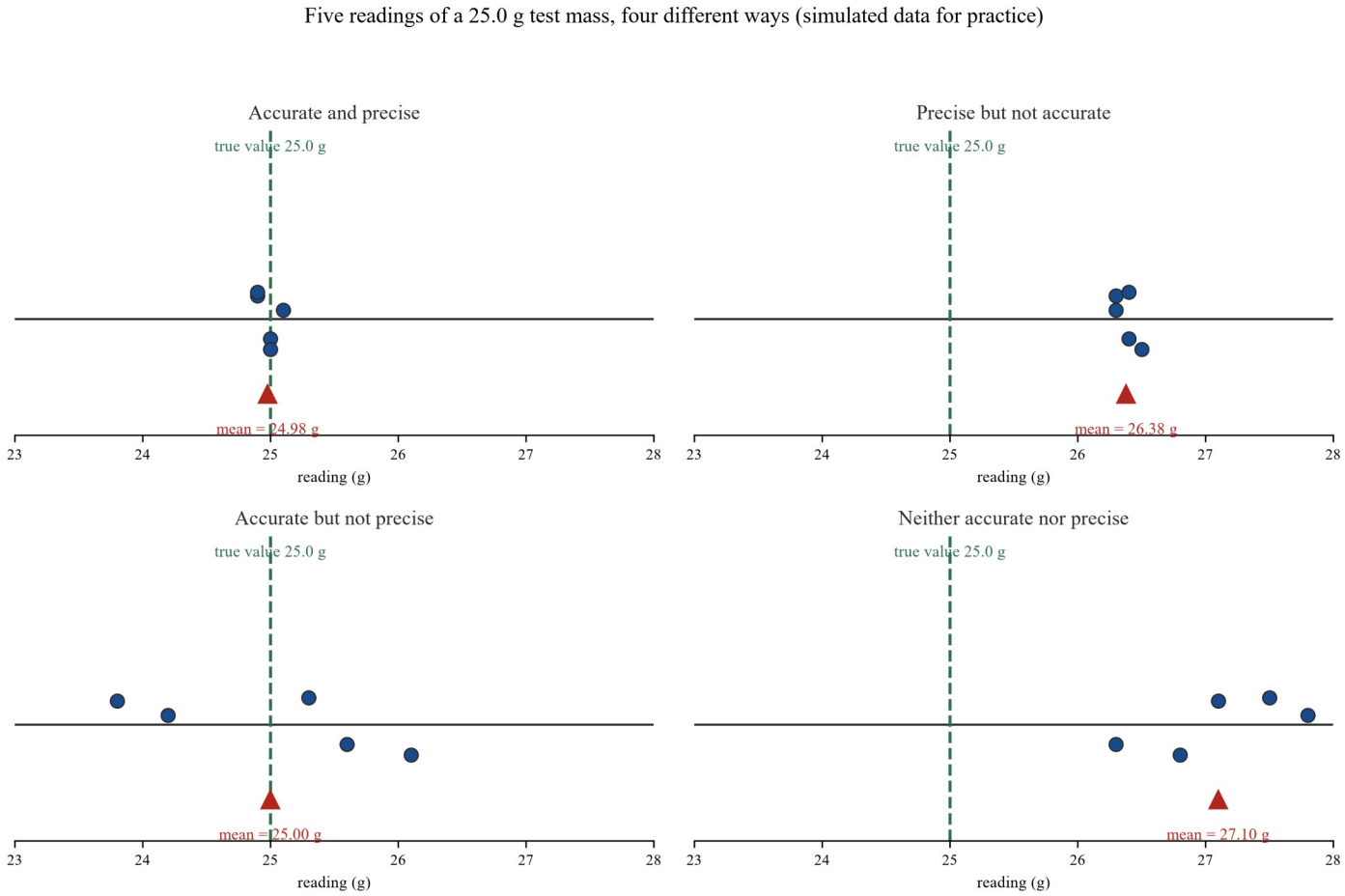
The inquiry sequence of this book as eight linked boxes: Question (2A), Hypothesis (2B), Plan (3A, 3B, 3C), Collect data, then Process and analyse (4A, 4B, 4C), EVALUATE (5A), Conclusion (5B) and Report (6A, 6B), with an amber arrow labelled "fix the method and repeat" leading from EVALUATE back up to Plan

Three words carried over from Subtopic 3A. This subtopic evaluates *validity* and *reproducibility*, so it works to the same definitions Subtopic 3A fixed:

- **valid** — an investigation is **valid** when it actually tests what it set out to test: the comparison is fair (only the factor being tested differs between the things compared), and the sampling and the observations genuinely answer the question asked.
- **reproducible** — an investigation is **reproducible** when someone else follows your written method and gets results that agree with yours.
- **bias** — **bias** is a push, built into the method, that shifts results one way. A mistake can push a result either way; bias pushes it the *same* way every time, so repeating a biased method does not fix it.

(Subtopic 3A used these words to *plan* good investigations. Here the same words are used to *judge* investigations after the data is in — your own, and other people's.)

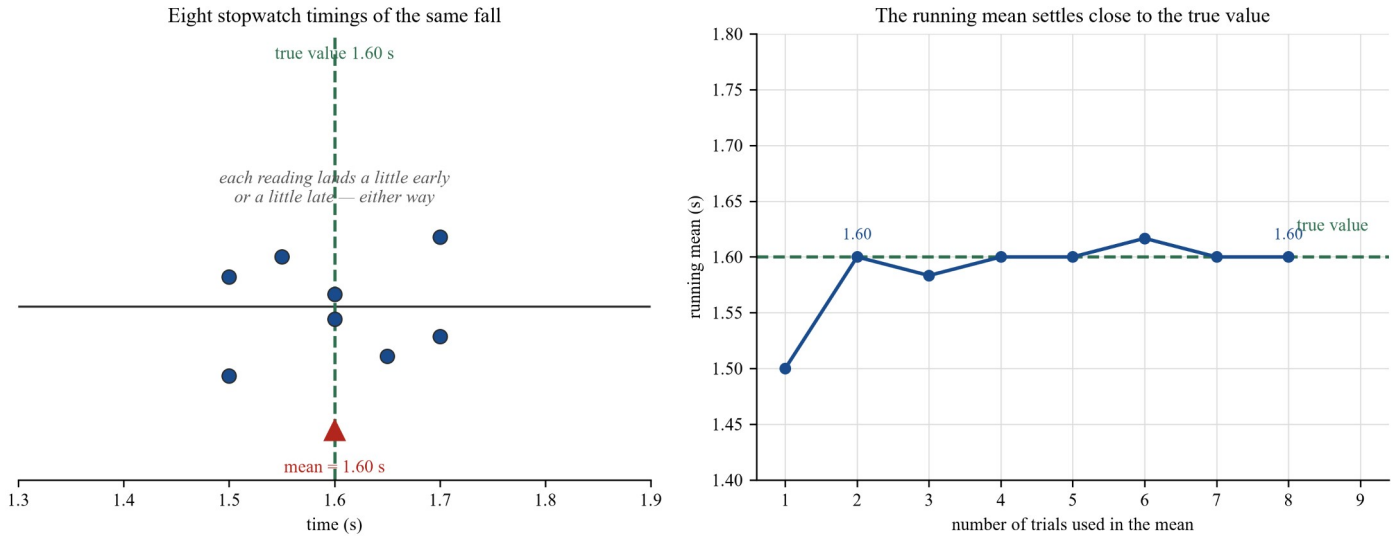
Accuracy: close to the true value. Subtopic 3C defined **precision** — how closely repeat readings agree with each other — and built the ± uncertainty that describes the spread of repeats. This subtopic adds the other half of quality: **accuracy**. A measurement (or the mean of a set of repeats) is **accurate** when it is close to the **true value**, the value a perfect measurement would give. Accuracy and precision are not the same thing, and the four panels below separate them.



Four dot-strip panels of five repeat mass readings each, against a green line marking the true value 25.0 g: "Accurate and precise" readings 24.9, 25.0, 25.1, 25.0, 24.9 g cluster tightly on the true value (mean 24.98 g); "Precise but not accurate" readings 26.3, 26.4, 26.3, 26.5, 26.4 g cluster tightly but 1.4 g high (mean 26.38 g); "Accurate but not precise" readings 23.8, 25.6, 24.2, 26.1, 25.3 g scatter widely but balance around the true value (mean 25.0 g); "Neither accurate nor precise" readings 27.1, 26.3, 27.8, 26.8, 27.5 g scatter widely and sit high (mean 27.1 g)

Read the panels across. **Accurate and precise**: the readings cluster tightly *and* on the true value — the mean, 24.98 g, is essentially 25.0 g. **Precise but not accurate**: the readings agree with each other (mean 26.38 g, a spread of only 0.2 g) but the whole cluster sits about 1.4 g high — precise, yet not close to the true value. **Accurate but not precise**: the readings scatter over 2.3 g, yet they scatter *both ways* around the true value, so the mean still lands on 25.0 g. **Neither**: scattered *and* shifted. The pattern to remember is that precision is about the *spread* of the readings, accuracy is about where they *sit* relative to the true value, and a precise result can be accurately wrong.

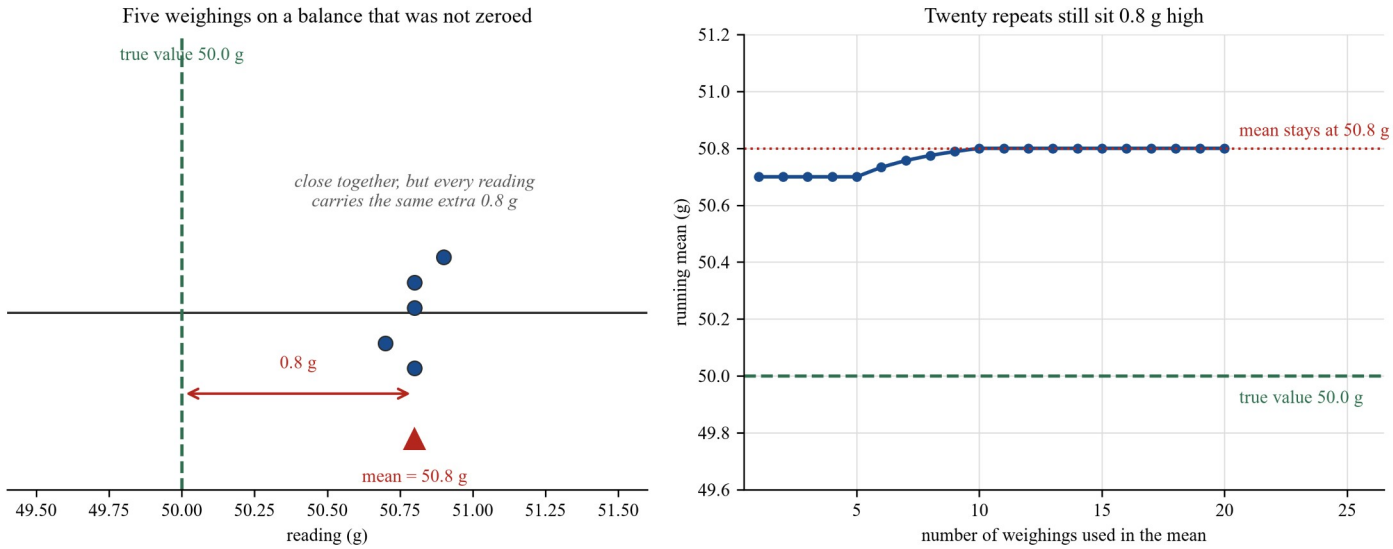
Random error — spread that goes both ways. No measurement is perfectly steady. When you time the same fall with a stopwatch, your thumb is a little early on one trial and a little late on the next, and the readings scatter above and below the true time by different amounts. That scatter is **random error**: the effect of unpredictable, ever-present variations (your reflexes, a breath of air, a hard-to-read scale) that push readings *both ways*. Random error shows up as spread, so it is what limits precision.



Two panels of eight stopwatch timings of the same fall, true time 1.60 s: left, the eight readings 1.5, 1.7, 1.55, 1.65, 1.6, 1.7, 1.5 and 1.6 s scattered both ways around the true value with mean 1.60 s; right, the running mean after 1, 2, 3 … 8 readings, settling towards 1.60 s as more repeats are averaged

The readings in the figure — 1.5, 1.7, 1.55, 1.65, 1.6, 1.7, 1.5 and 1.6 s — scatter from 1.5 to 1.7 s, but they scatter both ways, and the mean of the eight is exactly the true time, 1.60 s. That is the signature of random error, and it is also the good news: because the pushes cancel, **repeating and averaging reduces random error**. The right panel shows the running mean wobbling after two or three readings and settling onto 1.60 s as repeats are added. More repeats, tighter mean.

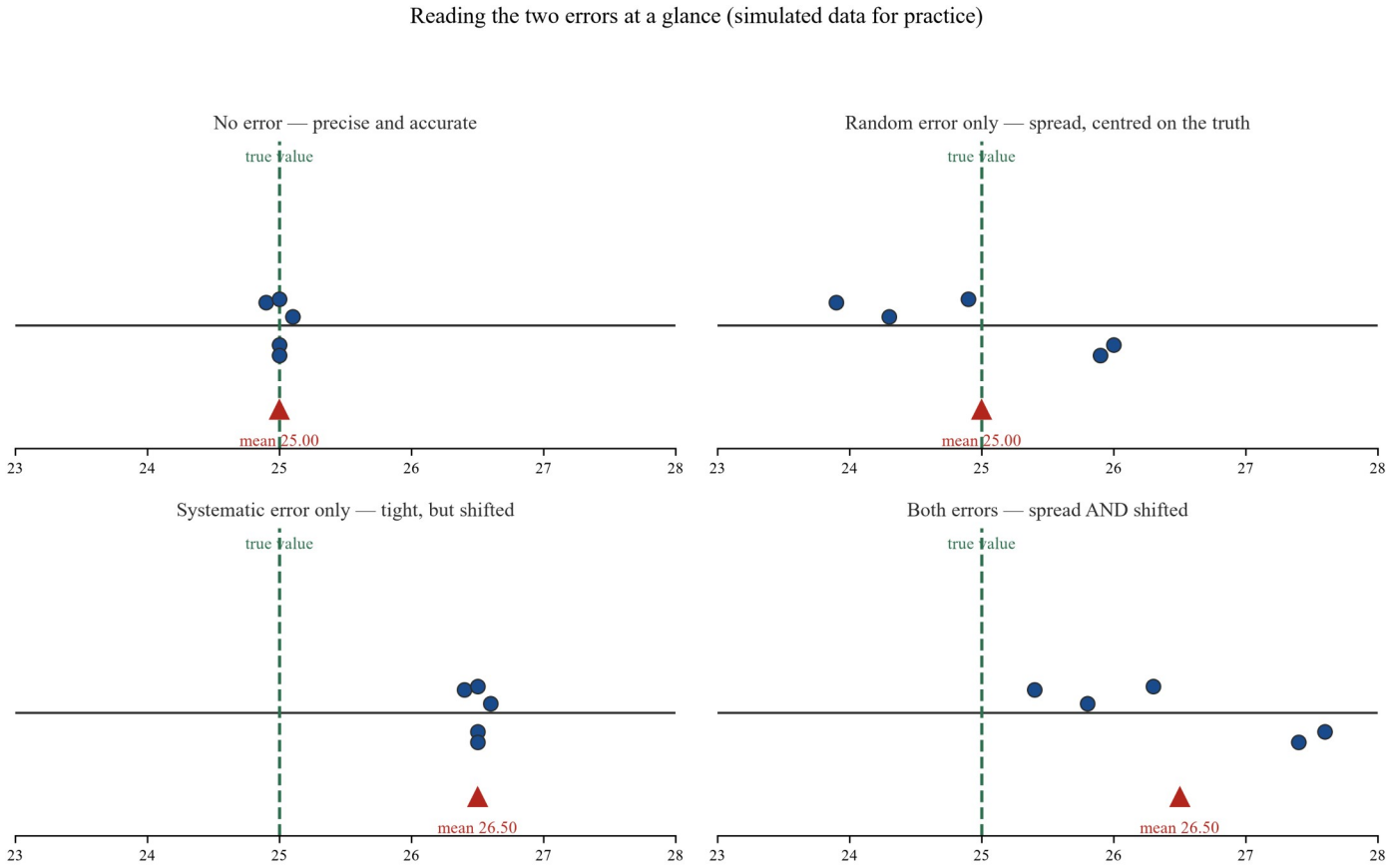
Systematic error — the same push every time. Now suppose the balance was never zeroed: with nothing on the pan it already reads 0.8 g. Every reading then carries the same extra 0.8 g — a 50.0 g mass reads 50.8 g, every single time. That steady shift is **systematic error**: a fault in the instrument or the method (a balance not zeroed, a ruler that expanded in the warm, a thermometer that reads 0.4 °C high in melting ice) that pushes *every* reading the *same* way. Systematic error shifts where the readings sit, so it is what limits accuracy.



Two panels of repeat mass readings of a 50.0 g mass on a balance that reads 0.8 g high: left, five readings 50.8, 50.7, 50.9, 50.8, 50.8 g clustered tightly 0.8 g above the green true-value line at 50.0 g; right, twenty repeats still clustered at mean 50.8 g — repeating has not moved the mean towards the true value

The cruel part of systematic error is in the right panel. The five readings are precise (mean 50.8 g, spread 0.2 g), and twenty repeats are *still* precise at 50.8 g — the mean never drifts towards 50.0 g, because **repeating does not reduce systematic error**. Every repeat carries the same push, so the mean carries it too. Systematic error is fixed only by fixing the fault: **calibrate** the instrument (Subtopic 3C checked a thermometer in melting ice for exactly this reason), check the zero before use, or change the method.

Reading the two errors from data. Put the two ideas side by side and data tells you which error you are looking at. Readings that scatter both ways around the true value show random error; a whole cluster shifted to one side shows systematic error; and the two can sit on the same data at once.

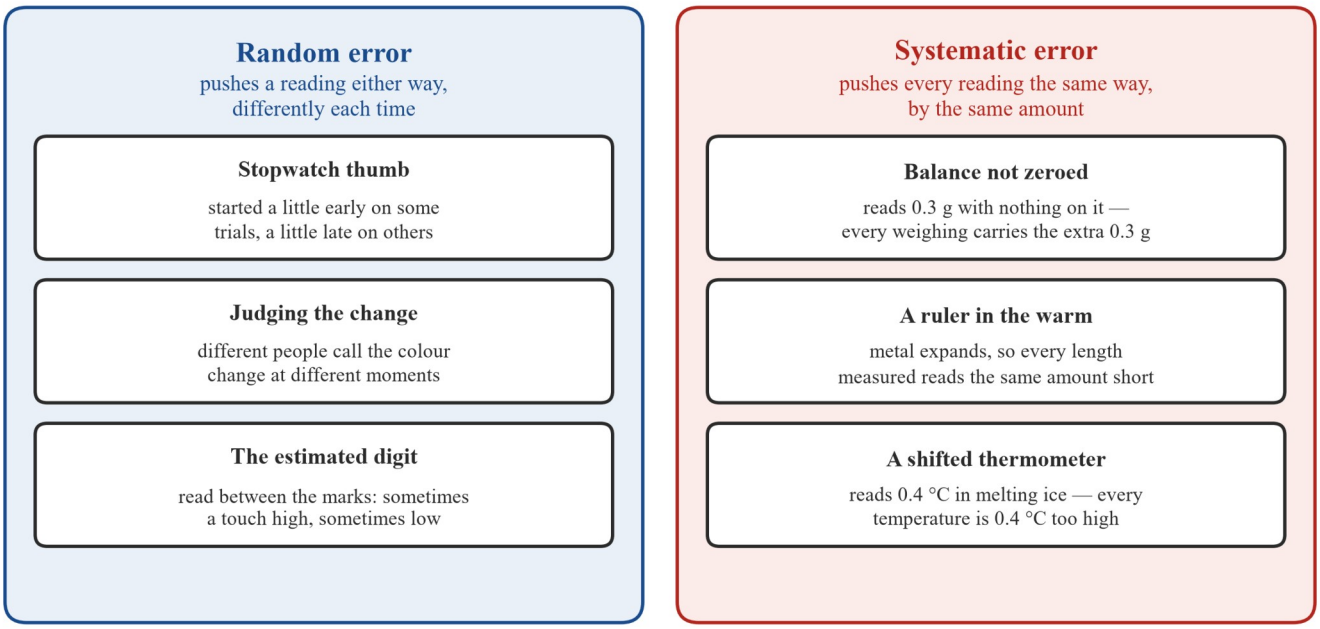


Four dot-strip panels of five repeat readings against the true value 25.0 g: "No error" readings 24.9, 25.0, 25.1, 25.0, 25.0 g tight on the true value (mean 25.0 g); "Random error only" readings 23.9, 26.0, 24.3, 25.9, 24.9 g scattered both ways with mean still 25.0 g; "Systematic error only" readings 26.4, 26.5, 26.6, 26.5, 26.5 g tight but shifted 1.5 g high (mean 26.5 g); "Both errors" readings 25.4, 27.6, 25.8, 27.4, 26.3 g scattered and shifted, mean 26.5 g

In the figure: scatter with the mean on the true value is random error only; a tight cluster off the true value is systematic error only; scatter *and* a shifted mean is both at once. The two questions to ask of any dataset are therefore: *how wide is the spread?* (random error) and *where does the mean sit relative to the true value, or to what it should be?* (systematic error).

Sorting the two errors. The rule is short. If the effect would change direction or size from reading to reading — a thumb early here, late there; an estimated digit read differently each time — it is random. If the effect would push every reading the same way — an unzeroed balance adding 0.3 g to everything; a warm metal ruler shrinking so every length reads short; a thermometer adding 0.4 °C to every temperature — it is systematic. Activity 1 practises the sort — this is the distinction elaboration `VC2S10I06.E02` names.

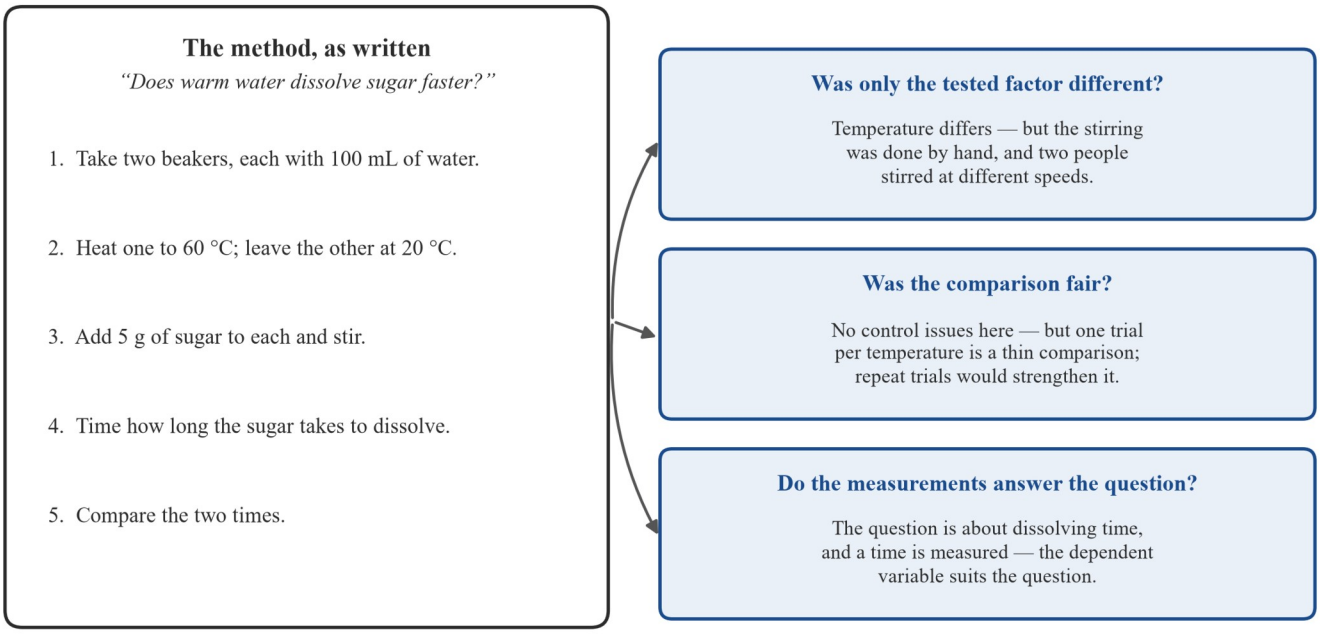
Sorting the sources of error



Six cards sorted into two columns: under "random error", the cards "Stopwatch thumb — early on some trials, late on others", "Judging the exact moment of a colour change" and "The estimated last digit of a scale reading"; under "systematic error", the cards "Balance not zeroed — reads 0.3 g with an empty pan", "A metal ruler used in the warm — it expands, so lengths read short" and "A thermometer that reads 0.4 °C high in melting ice"

Evaluating a method for validity. A method is evaluated against the question it was meant to answer, using three checks: *was only the tested factor different? was the comparison fair and repeated enough to trust? do the measurements actually answer the question?* The checks are easier to run on a real method than on a definition, so here is one.

Evaluating the validity of a method: three questions



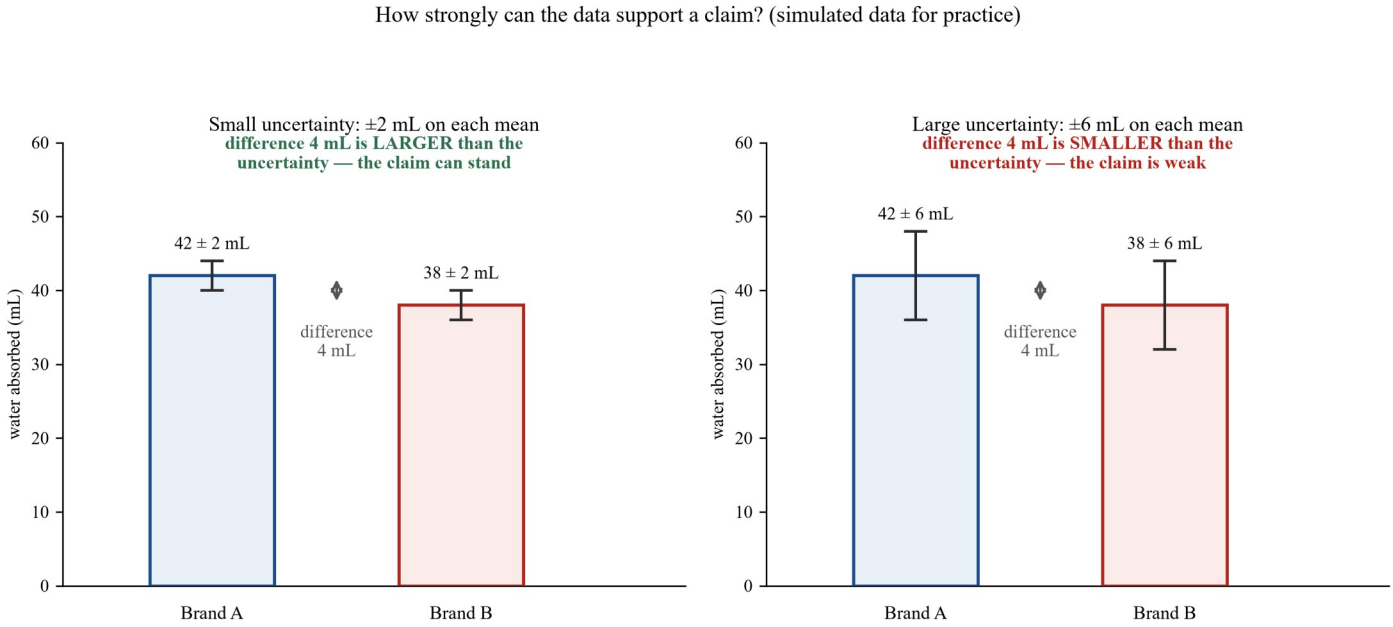
A method card and three evaluation checks. Method: "Does warm water dissolve sugar faster? Two beakers each hold 100 mL of water, one at 60 °C and one at 20 °C. Add 5 g of sugar to each and stir by hand. Time how long each takes to dissolve." Check 1, "Was only the tested factor different?", flagged as a flaw because stirring by hand may

have stirred one beaker harder; Check 2, "Was the comparison fair?", flagged as thin because there was only one trial per temperature; Check 3, "Do the measurements answer the question?", passed because the time to dissolve is exactly what the question asks about

Run the three checks. The tested factor is temperature, and the volumes and sugar mass are controlled — but *stirring by hand* means one beaker may have been stirred harder, and stirred harder dissolves faster: a second factor may differ, so the first check raises a flaw. One trial per temperature is a thin comparison — a single pair of beakers could be an unlucky pair — so the second check wants repeats. The third check passes: timing how long the sugar takes to dissolve is exactly what "faster" means. An investigation is valid only to the degree that checks like these come out clean, and each flaw found names the fix: stir both beakers the same way (or not at all), and repeat the trials.

Evaluating reproducibility. Reproducibility is tested by another group, not by you: someone follows your *written* method and compares results. That makes the written method the key document — amounts, times, and the rules for what counts all stated, so that nothing lives only in your head. When a repeat does not agree, the method, the instruments (calibration again) and the sample are the first places to look, before anyone accuses the data.

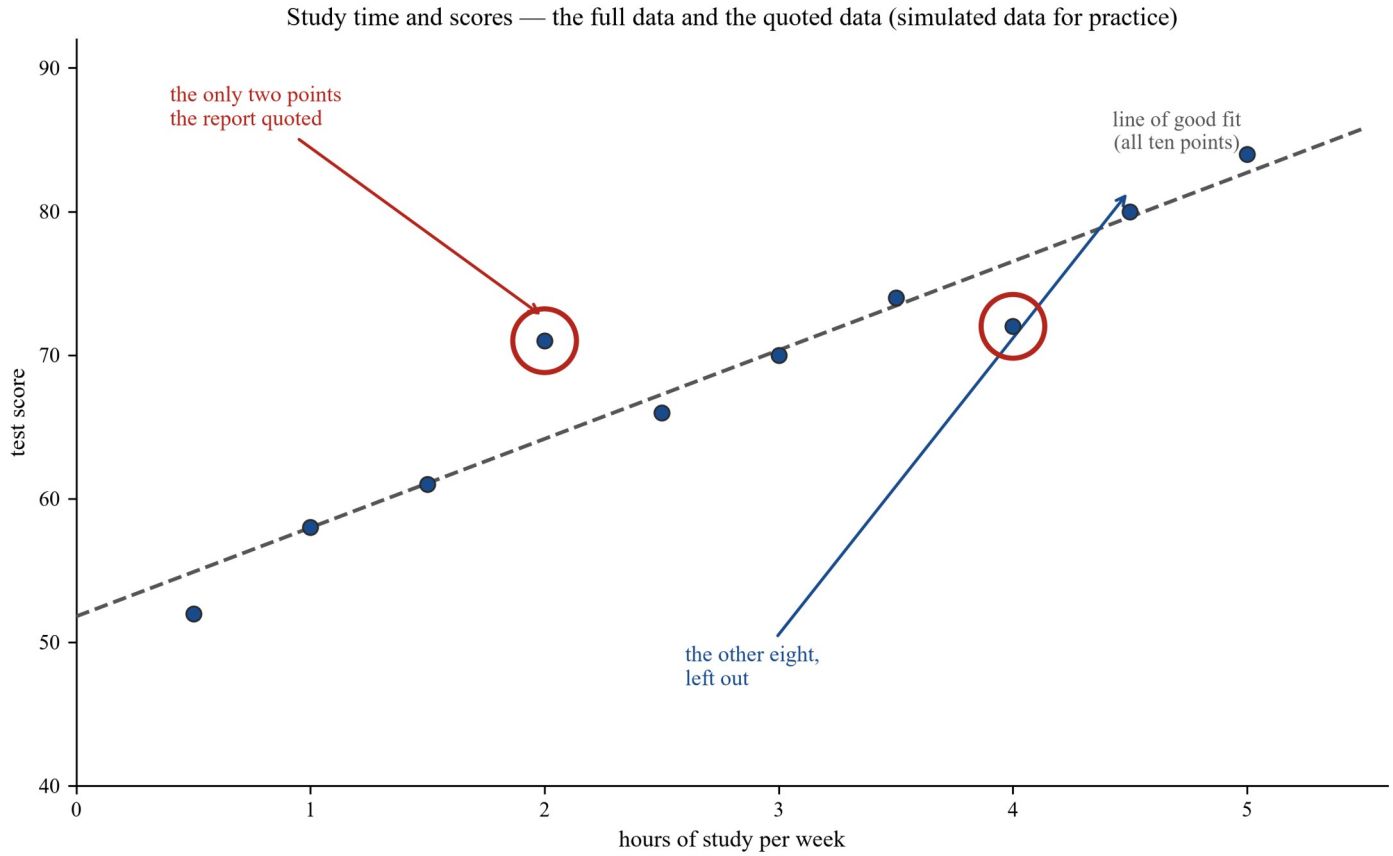
How strong is the conclusion? (elaboration `VC2S10I06.E01`). A conclusion is only as strong as the evidence under it, and elaboration `VC2S10I06.E01` is about judging exactly that strength. The single most useful test is to compare the *difference* a claim rests on with the *uncertainty* of the data (the ± spread of repeats from Subtopic 3C). If the difference is bigger than the uncertainty, the claim can stand; if the uncertainty swallows the difference, the claim is weak — not necessarily wrong, but not proven by this data.



Two panels comparing water absorbed by two paper-towel brands, Brand A 42.0 mL and Brand B 38.0 mL, a difference of 4 mL. Left panel: uncertainty bars of ± 2 mL on each brand do not overlap, and the verdict reads "difference 4 mL, uncertainty ± 2 mL — the claim can stand". Right panel: uncertainty bars of ± 6 mL overlap across the two brands, and the verdict reads "difference 4 mL, uncertainty ± 6 mL — the claim is weak"

The figure runs the test twice on the same means, 42.0 mL against 38.0 mL. With repeats giving ± 2 mL, the uncertainty bars do not overlap: the 4 mL difference is twice the uncertainty, and "Brand A soaks up more" can stand. With ± 6 mL, the bars overlap heavily: the brands might differ, or the 4 mL might just be spread, and the same conclusion is weak. Same data, same difference — different strength of conclusion. Size of the difference against size of the uncertainty, number of repeats, and the checks of validity above are what `VC2S10I06.E01` means by the strength a conclusion can carry.

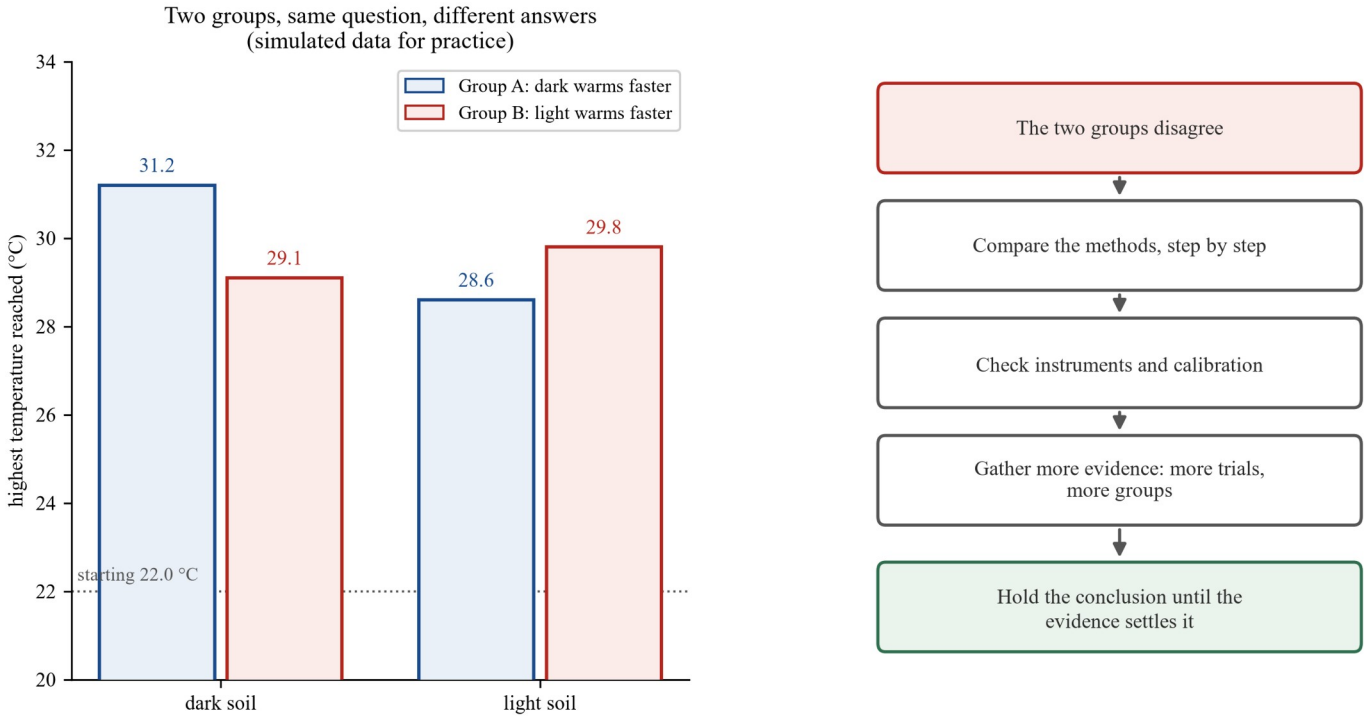
Bias in what gets shown. Bias is not always in the measuring; it can be in the *choosing* of what evidence gets shown. A report that quotes the two points that support its story and leaves the other eight in the drawer has not invented data — it has selected it, and selection is a bias that influences conclusions as much as a shifted instrument does.



A scatterplot of ten students' study time against test score (simulated data for practice): hours 0.5 to 5.0 in steps of 0.5 with scores 52, 58, 61, 71, 66, 70, 74, 72, 80 and 84. A dashed grey line of good fit rises steadily through all ten points. The two points (2, 71) and (4, 72) are ringed in red as "the only two points the report quoted", and a blue arrow labels the remaining eight as "the other eight, left out"

Look at what the two quoted points "show": from 2 hours (score 71) to 4 hours (score 72), almost no gain — "extra study makes no difference". Now add the eight points the report left out: scores climb from 52 at half an hour to 84 at five, and the line of good fit through *all ten* points rises steadily. The quoted pair sat in the middle of the trend and hid it. Evaluating a claim therefore always includes the question: *is this all the evidence, or all the evidence that was shown?*

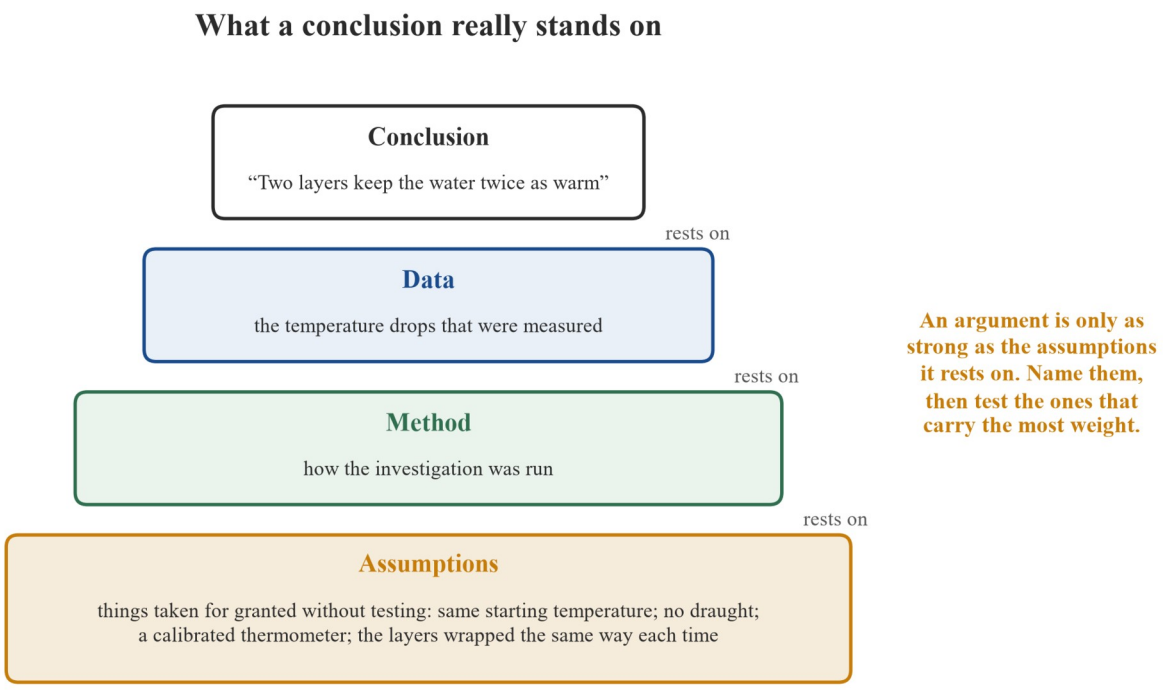
Conflicting evidence. Sometimes the evidence disagrees openly — two groups run the same-style investigation and get different answers. Conflicting evidence does not mean someone cheated; it is ordinary in science, and the curriculum expects you to work with it, not around it.



Two panels of soil temperature after one hour in the sun, starting at 22.0 °C: Group A measured dark soil 31.2 °C and light soil 28.6 °C; Group B measured dark soil 29.1 °C and light soil 29.8 °C. Below, a five-step response: 1 the groups disagree; 2 compare the methods; 3 check the instruments and calibration; 4 gather more evidence; 5 hold the conclusion until the conflict is explained

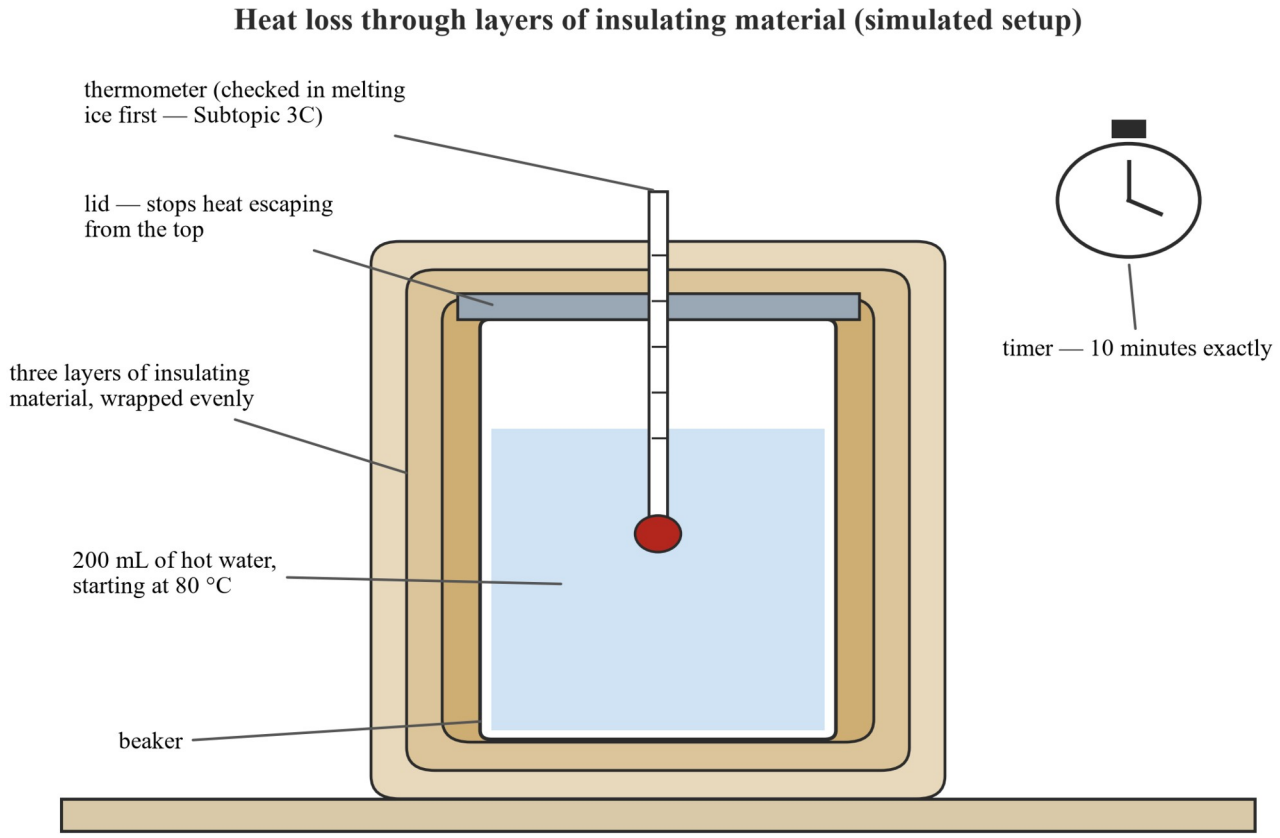
Group A found dark soil warms clearly more (31.2 °C against 28.6 °C); Group B found almost no difference, and in the other direction (29.1 against 29.8 °C). The figure's five steps are the disciplined response: name the disagreement; compare the methods for a difference (different soil moisture? different depth of the thermometer?); check the instruments and their calibration; gather more evidence; and *hold* the conclusion until the conflict is explained. A conclusion drawn while conflicting evidence stands unexamined is a weak conclusion, whatever it claims.

Assumptions — what the argument rests on (elaboration `VC2S10I06.E03`). Every investigation stands on things taken for granted — an **assumption** is a fact or premise accepted as true without being checked in this investigation. Assumptions are not cheating; every method has them, because you cannot check everything at once. What evaluation demands is that they are *named*, and that the ones carrying the most weight are tested.

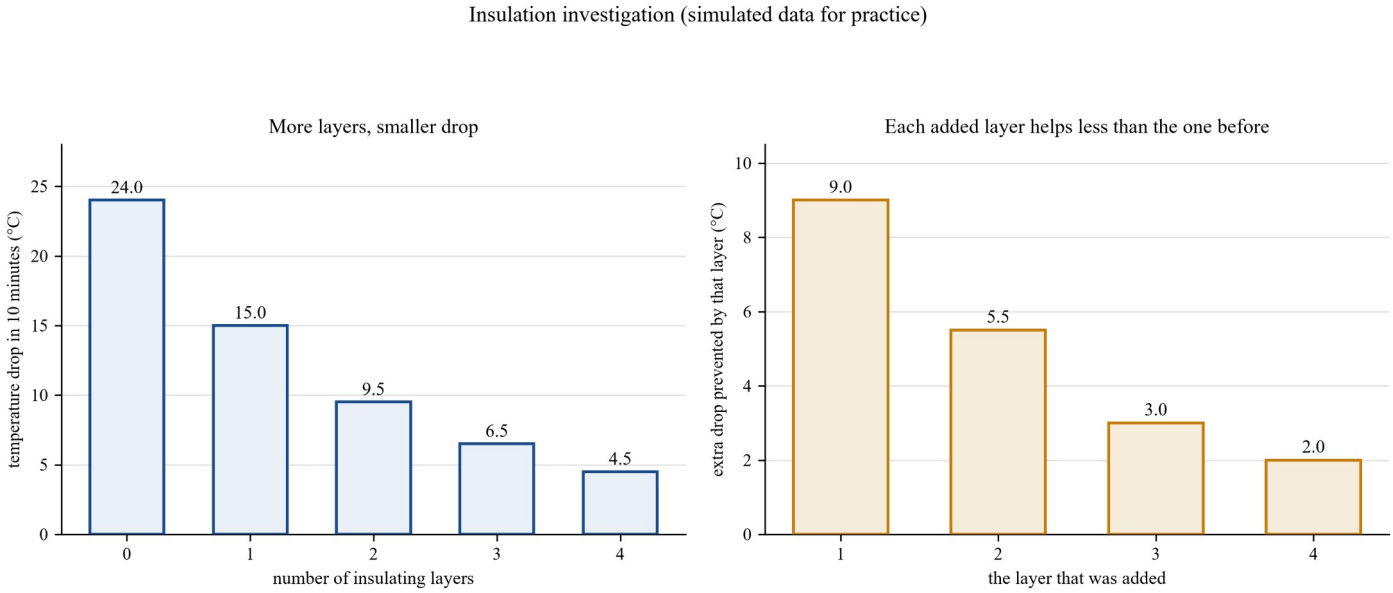


A stack of four boxes, each resting on the one below: at the top "Conclusion — 'Two layers keep the water twice as warm'", resting on "Data", resting on "Method", resting on the widest box "Assumptions — same starting temperature; no draught; calibrated thermometer; layers wrapped the same way each time". A side note reads: an argument is only as strong as the assumptions it rests on — name them, then test the ones that carry the most weight

The insulation investigation — evaluating a real method (elaboration `VC2S10I06.E03`). Elaboration `VC2S10I06.E03` evaluates methods and conclusions about heat moving through layers of an insulating material. Here is the method to evaluate. This subtopic does not ask *why* the layers slow the cooling — that mechanism is heat transfer, taught later in the book. Evaluation works on the method and the data alone, and it needs nothing more.



The apparatus: a beaker of 200 mL of hot water starting at 80 °C with a lid and a thermometer (checked in melting ice first), wrapped evenly in three layers of insulating material, timed for 10 minutes exactly, with the room temperature held at 20 °C

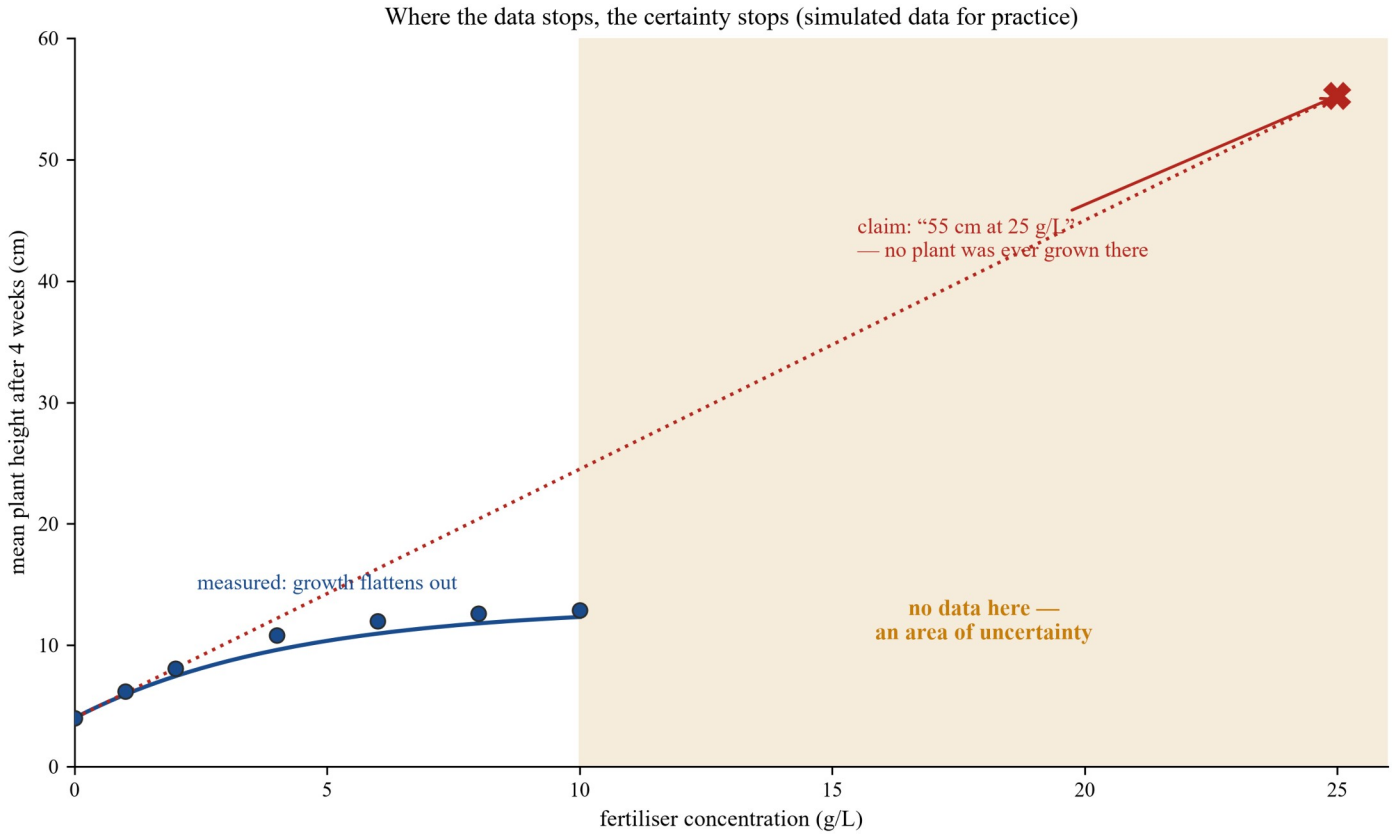


Simulated data for practice: temperature drop in 10 minutes for 0, 1, 2, 3 and 4 layers of insulating material — drops of 24.0, 15.0, 9.5, 6.5 and 4.5 °C, a falling curve that flattens; brackets between the points show what each added layer saves, 9.0, 5.5, 3.0 and 2.0 °C, each added layer helping less than the one before

First, the assumptions. The method assumes the water started at the same temperature every run; that no draught cooled one run more than another; that the thermometer read truly (hence the melting-ice check); and that the layers were wrapped the same way each time. None of those is checked *by* the method — they are assumed *for* it. Each is widely accepted as reasonable in a classroom (the starting temperature can be read, the thermometer can be checked), and each would, if false, shift the results: a draught on the zero-layer run would make the layers look better than they are. Naming them turns “the result is the result” into “the result stands on these four things”.

Second, the data. The drop falls from 24.0 °C with no layers to 4.5 °C with four — the layers clearly reduce the cooling. But the brackets tell the fuller story: each added layer saves 9.0, then 5.5, then 3.0, then 2.0 °C — **each added layer helps less than the one before**. A conclusion that "every layer helps the same amount" would not survive this data; a conclusion that "more layers reduce the cooling, with a shrinking return" would. That comparison — claim against what the data actually shows — is the evaluation `VC2S10I06.E03` asks for, and it is also `VC2S10I06.E01` again: the strength of the conclusion is set by the data under it.

Areas of uncertainty. Data covers the range that was actually measured, and a conclusion is safest inside that range. Beyond it lies an **area of uncertainty** — a range the data does not cover, where any claim rests on assuming the pattern continues. Extending a pattern a little way past the last point is sometimes reasonable; extending it far, or across a range where the pattern was already changing shape, is a claim without evidence under it.

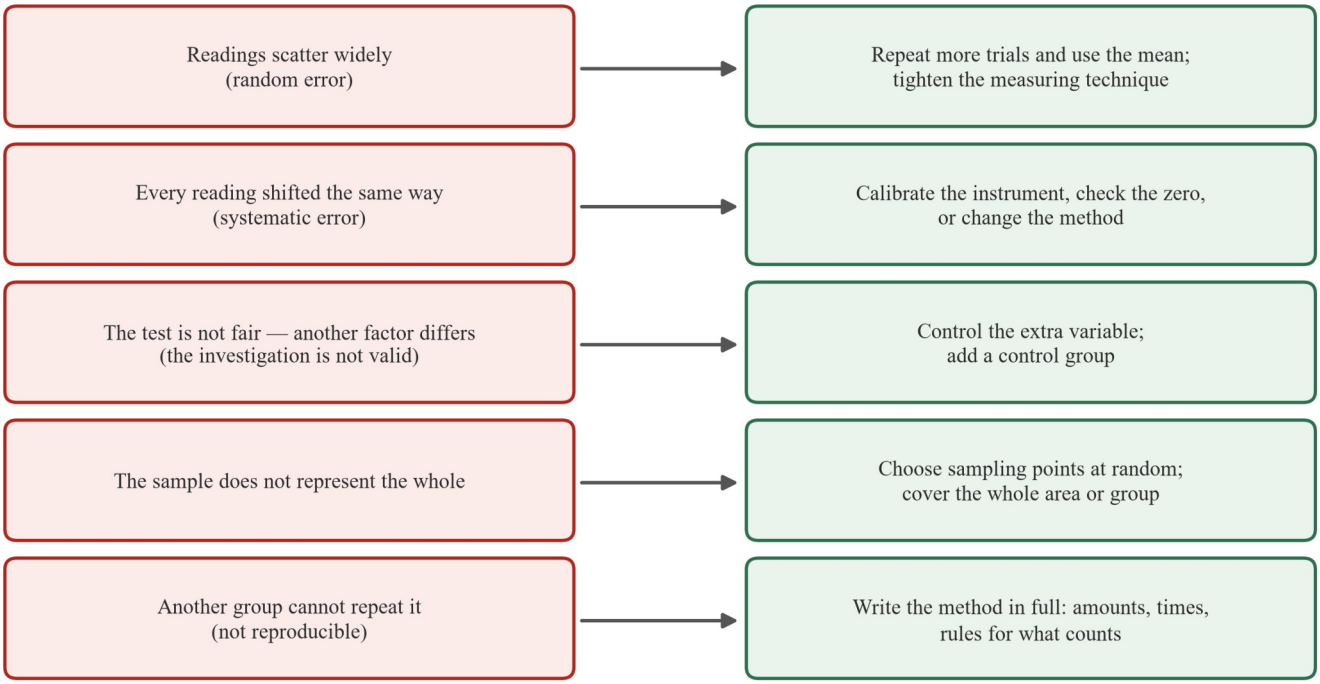


A scatterplot of mean plant height after 4 weeks against fertiliser concentration (simulated data for practice): concentrations 0, 1, 2, 4, 6, 8 and 10 g/L give heights 4.0, 6.2, 8.1, 10.8, 12.0, 12.6 and 12.9 cm — a rising curve that flattens towards a plateau, drawn as a solid blue line through the measured points. A red dotted line extends the early steep rise far beyond the data to a large X at 25 g/L and about 55 cm, labelled "claim: 55 cm at 25 g/L — no plant was ever grown there". The region beyond 10 g/L is shaded amber and labelled "no data here — an area of uncertainty"

The measured data runs from 0 to 10 g/L, and inside that range it is clear: growth rises, then flattens towards about 12.9 cm. The claim reads the early rise as a straight line and extends it to 25 g/L — "55 cm" — a concentration where no plant was ever grown. Two things fail. The 10 to 25 g/L range is an area of uncertainty: nothing was measured there, so nothing supports any number at all. And even the shape of the measured curve argues against the claim: growth was *already flattening* by 8 g/L, so assuming the early steep rise continues is assuming a pattern the data has stopped showing. Where the data stops, the certainty stops.

Improving the quality of data. The achievement standard for this subtopic ends where evaluation should end: with the fix. Every problem evaluation finds has a matching way to improve the data, and the pairings are worth knowing cold.

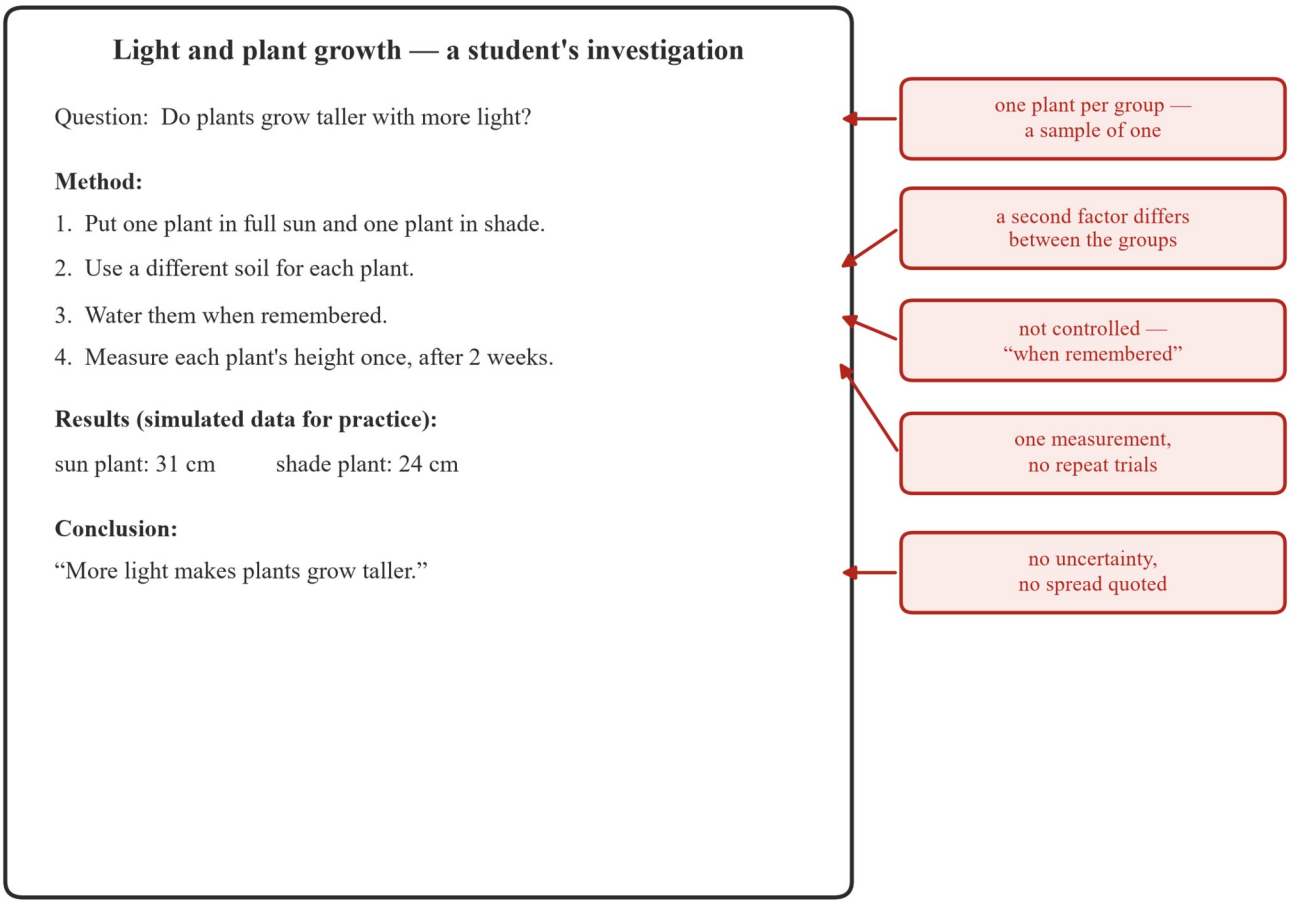
Ways to improve the quality of data — match the fix to the problem



Five problem-and-fix pairs. "Readings scatter widely (random error)" pairs with "Repeat more trials and use the mean; tighten the measuring technique". "Every reading shifted the same way (systematic error)" pairs with "Calibrate the instrument, check the zero, or change the method". "The test is not fair — another factor differs (the investigation is not valid)" pairs with "Control the extra variable; add a control group". "The sample does not represent the whole" pairs with "Choose sampling points at random; cover the whole area or group". "Another group cannot repeat it (not reproducible)" pairs with "Write the method in full: amounts, times, rules for what counts"

And that closes the loop drawn in the first figure: evaluation judges the method, the data and the claim — then names the fix, feeds it back into the plan, and the next round of data is better because of it.

A whole investigation to evaluate. The report below brings everything in this subtopic together — run the checks above on it now, and return to it in the activities and questions that follow.



Evaluate this investigation: what stands, what does not, and what would fix it?

The student report to evaluate: "Light and plant growth — a student's investigation". Question: Do plants grow taller with more light? Method: 1. Put one plant in full sun and one plant in shade. 2. Use a different soil for each plant. 3. Water them when remembered. 4. Measure each plant's height once, after 2 weeks. Results (simulated data for practice): sun plant 31 cm, shade plant 24 cm. Conclusion: "More light makes plants grow taller." Five red boxes flag: one plant per group — a sample of one; a second factor differs between the groups; not controlled — "when remembered"; one measurement, no repeat trials; no uncertainty, no spread quoted. Caption: Evaluate this investigation — what stands, what does not, and what would fix it?

Activities

Activity 1 — Sort the errors. Work in pairs. Copy the six cards of the error-sort figure onto cardstock (or write your own six: three effects that would push readings both ways, three that would push every reading the same way), shuffle them, and sort them into *random* and *systematic*, saying the reason aloud each time: "it changes direction, so random" / "same push every time, so systematic". Then, for one card from each column, state the fix: what reduces or removes that error? Check your sorts against the figure and against each other — any card you disagree on is worth a whole class discussion.

Activity 2 — Audit the insulation investigation (elaboration `VC2S10I06.E03`). Using the apparatus and data figures of the theory: (1) list every assumption the method rests on, including ones the figure does not name (think about the lid, the bench, the person reading the thermometer); (2) rank your assumptions by how much the conclusion leans on them; (3) for your top two, write one check each that would test whether the assumption holds; (4) state, from the data, whether "more layers reduce the cooling" and whether "each extra layer saves the same amount" are supported, with the numbers that decide it. No heat-transfer theory is needed or wanted — the audit is about method, assumptions and data only.

Activity 3 — Judge the report, then fix it. Use the student report figure at the end of the theory (Question 14 returns to it). In pairs, play the part of a journal referee: write three things about the investigation that stand, three that do not,

and the single change that would most improve the quality of the data, with one sentence of reason for each. Then rewrite the method and the conclusion the way you would accept them.

Worked examples

Datasets in this section are simulated data for practice.

Example 1 — scaffolded

Two datasets are described. Classify the main error in each as random or systematic, and justify each classification with the signature of that error.

(i) Eight stopwatch times for the same fall: 1.5, 1.7, 1.55, 1.65, 1.6, 1.7, 1.5, 1.6 s.

(ii) A thermometer reads 0.4 °C when held in melting ice (it should read 0.0 °C). It is then used to record a cooling curve: 20.4, 24.4, 28.4 °C where the true temperatures are 20.0, 24.0, 28.0 °C.

Working	Comment
(i) The readings scatter from 1.5 to 1.7 s — above the true time on some trials, below it on others	Random error scatters readings *both ways* by different amounts.
The mean of the eight, 1.60 s, sits on the true time	Both-way scatter cancels in the mean — the signature of random error.
(ii) Every reading is 0.4 °C above the true value: 20.4 vs 20.0, 24.4 vs 24.0, 28.4 vs 28.0	The same push, every reading, same direction.
Repeating would not move the mean: every repeat carries the same +0.4 °C	Systematic error is not reduced by repeats — the signature of systematic error.

Dataset (i) shows **random error** (both-way scatter, mean on the true value); dataset (ii) shows **systematic error** (every reading shifted the same way).

Example 2 — scaffolded

A class compares two paper-towel brands, five repeats each: Brand A soaks a mean of 42.0 mL, Brand B a mean of 38.0 mL. With the repeats giving an uncertainty of ± 2 mL, the class concludes “Brand A soaks up more water than Brand B”.

(a) State the difference the conclusion rests on.

(b) Judge the strength of the conclusion with an uncertainty of ± 2 mL.

(c) Judge it again if a sloppier technique had given ± 6 mL.

Working	Comment
(a) 42.0 − 38.0 = 4.0 mL	The conclusion rests on this difference.
(b) Difference 4.0 mL against uncertainty ± 2 mL: the difference is twice the uncertainty; the uncertainty bars do not overlap	Difference bigger than the uncertainty → the claim can stand.
(c) Difference 4.0 mL against uncertainty ± 6 mL: the uncertainty is bigger than the difference; the bars overlap	Uncertainty swallows the difference → the same conclusion is weak.

With ± 2 mL the conclusion can stand; with ± 6 mL the 4.0 mL difference could just be spread, and the conclusion is weak. Same data, different strength — the strength lives in the size of the difference *relative to* the uncertainty.

Example 3 — unscaffolded

Evaluate the validity of the sugar method of the theory figure (“Does warm water dissolve sugar faster?”, two 100 mL beakers at 60 °C and 20 °C, 5 g of sugar each, stirred by hand, timed to dissolve), using the three validity checks.

The third check passes: the time taken to dissolve is exactly what "faster" asks for, so the measurement answers the question. The first check finds the flaw that matters most: the tested factor is temperature, but stirring *by hand* means a second factor — how hard each beaker was stirred — may also differ, and stirring speeds up dissolving; if the warm beaker happened to be stirred harder, its faster dissolving proves nothing about temperature. The second check finds the comparison thin: one trial per temperature is a single pair of beakers, which could be an unlucky pair, so the result has no repeats behind it. The method is therefore **not valid as written**: the fix is to stir both beakers the same way (or not at all) and to repeat the trials, after which the method would test what it set out to test.

Example 4 — unscaffolded

A fertiliser company's report quotes the early, steep part of the theory's plant data (growth rising from 4.0 cm at 0 g/L) and claims: "at 25 g/L, plants will grow about 55 cm tall". Evaluate the claim against the full dataset, which runs to 10 g/L with heights 10.8, 12.0, 12.6 and 12.9 cm at 4, 6, 8 and 10 g/L.

The measured data stops at 10 g/L, so 25 g/L sits in an area of uncertainty — no plant was grown there, and no measurement supports any height at all at that concentration. The claim also assumes the early steep rise continues as a straight line, but the measured curve had already flattened: from 8 to 10 g/L growth rose only 12.6 → 12.9 cm, a gain of 0.3 cm, not the roughly 2 cm per g/L the early rise showed. A pattern that has flattened inside the measured range gives no licence to extend the steep line far beyond it. The claim of about 55 cm at 25 g/L is therefore **not supported**: it rests on a pattern the data stopped showing, extended into a range the data never covered.

Practice questions

Scaffolded questions

Question 1 Use the inquiry-sequence figure. (a) Name the step that comes immediately before EVALUATE in the sequence. (b) Name the step the amber "fix the method and repeat" arrow leads back to. (c) Explain in one sentence why evaluation is placed before the conclusion is accepted.

Question 2 Use the four-panel accuracy–precision figure (true value 25.0 g). (a) Which panel is precise but not accurate? Quote its mean. (b) Which panel is accurate but not precise? Explain how a scattered set of readings can still be accurate. (c) Which error — random or systematic — best explains the "precise but not accurate" panel? Give a reason.

Question 3 Use the random-error figure (eight stopwatch readings, true time 1.60 s). (a) Calculate the mean of the eight readings. (b) Describe how the readings sit around the true value. (c) State the step that reduces the effect of this kind of error, and what the right-hand panel shows about it.

Question 4 Use the error-sort figure of six cards. (a) Place the "stopwatch thumb" and the "balance not zeroed" cards in their columns, with the one-line reason for each. (b) A seventh card reads: "a draught blows across the balance pan, differently each time". Which column does it join, and why? (c) For the "thermometer reads 0.4 °C high" card, state the fix.

Question 5 Use the sugar-method validity figure. (a) Which validity check fails because of hand stirring, and what is the extra factor it lets in? (b) Which check complains about one trial per temperature? (c) Which check passes, and why?

Question 6 Use the strength-of-conclusion figure (Brand A 42.0 mL, Brand B 38.0 mL). (a) State the difference between the brands. (b) Explain why the conclusion "Brand A soaks up more" can stand with an uncertainty of ± 2 mL. (c) Explain why the same conclusion is weak with an uncertainty of ± 6 mL.

Unscaffolded questions

Question 7 A balance reads 0.8 g with an empty pan. Five readings of a 50.0 g mass on it are 50.8, 50.7, 50.9, 50.8 and 50.8 g. (a) Calculate the mean of the five readings. (b) Name the type of error shown, with the evidence from the data. (c) Explain why twenty more repeats would *not* bring the mean to 50.0 g. (d) State two ways to fix the fault.

Question 8 A thermometer reads 0.4 °C in melting ice instead of 0.0 °C, and is then used to record a cooling curve every 2 minutes. (a) Name the type of error the fault produces. (b) State the effect on each recorded temperature. (c)

State the effect on the temperature *fall* between any two readings, with a reason. (d) Does this fault make the cooling-curve conclusion weaker, or leave it much as it was? Justify.

Question 9 Use the conflicting-evidence figure (Group A: dark soil 31.2 °C, light soil 28.6 °C; Group B: dark 29.1 °C, light 29.8 °C; both starting at 22.0 °C). (a) State, in one sentence each, what Group A and Group B found. (b) Suggest two differences in method or instruments that could produce the disagreement. (c) State what the class should do with the conclusion "dark soil warms more" while the conflict stands unexplained.

Question 10 Use the insulation assumptions figure (elaboration `VC2S10I06.E03`). (a) List two assumptions the investigation rests on. (b) Choose one of them and describe a check that would test whether it holds. (c) Explain what happens to the conclusion if that assumption turns out false.

Question 11 Use the insulation data figure (drops of 24.0, 15.0, 9.5, 6.5 and 4.5 °C for 0 to 4 layers). (a) Describe the pattern in the temperature drop as layers are added. (b) Calculate what each added layer saves (the four brackets). (c) Judge two conclusions against the data: "more layers reduce the cooling" and "each extra layer saves the same amount". Which is supported, which is not, and by what numbers?

Question 12 Use the cherry-picked scatterplot (ten study-time/score points; the report quoted only (2, 71) and (4, 72)). (a) State what the two quoted points, taken alone, appear to show. (b) Describe what the line of good fit through all ten points shows instead. (c) Name the bias at work, and state the question an evaluator should always ask when a report quotes only part of a dataset.

Question 13 Use the areas-of-uncertainty figure (measured 0–10 g/L; growth flattening at 12.9 cm; claim of 55 cm at 25 g/L). (a) State the highest concentration actually measured. (b) Explain why the 25 g/L claim sits in an area of uncertainty. (c) Explain why the *shape* of the measured data argues against the claim even inside the measured range.

Question 14 Use the student report figure ("Light and plant growth": one plant in full sun, one in shade, different soils, watered when remembered, one height measurement after 2 weeks; results sun 31 cm, shade 24 cm; conclusion "More light makes plants grow taller."). (a) Identify the two flaws that make the comparison not valid. (b) Identify the flaw that makes the investigation not reproducible. (c) Identify the flaw in the results (one measurement per plant), and name the kind of error repeats would address. (d) Rewrite the conclusion in language its evidence can carry.

Question 15 Group X writes up its bounce-height method in full. Group Y follows the written method and gets means that disagree with Group X's by several centimetres. (a) State what "reproducible" means. (b) Give two things Group Y should check against the written method before doubting the data. (c) If the methods truly match and both rulers were checked, state what the groups should do next to resolve the disagreement.

Question 16 For each scenario, name the problem type (random error / systematic error / not valid / unrepresentative sample / not reproducible) and give the matching fix from the improve figure. (a) Repeat readings of the same mass scatter over 1.4 g. (b) A survey of creek health samples only the spot beside the footbridge. (c) Every temperature in an experiment is 0.4 °C high because the thermometer was never checked. (d) Another class cannot repeat your investigation because your method says "heat for a while".

Answers

Question 1 (a) Process and analyse (4A · 4B · 4C); (b) Plan (3A · 3B · 3C); (c) the conclusion should rest on a method that has been judged — evaluation finds the flaws so the plan can fix them before the claim is trusted. **Question 2** (a) "Precise but not accurate", mean 26.38 g; (b) "Accurate but not precise" — the readings scatter both ways around 25.0 g, so the mean still lands on the true value; (c) systematic error — the whole tight cluster is shifted one way (about 1.4 g high). **Question 3** (a) 1.60 s; (b) scattered above and below it, both ways; (c) repeat and average — the running mean settles onto 1.60 s as repeats are added. **Question 4** (a) stopwatch thumb → random (early or late, changes direction); balance not zeroed → systematic (adds 0.3 g to every reading); (b) random — the draught pushes differently each time; (c) calibrate it against a known temperature (check it in melting ice) or replace it. **Question 5** (a) check 1, "was only the tested factor different?" — it lets in stirring effort as a second factor; (b) check 2, "was the comparison fair?" (one trial per temperature); (c) check 3 — the time to dissolve is exactly what "faster" means. **Question 6** (a) 42.0 − 38.0 = 4.0 mL; (b) the difference (4.0 mL) is twice the uncertainty (± 2 mL), and the bars do not overlap, so the difference is real against the spread; (c) the uncertainty (± 6 mL) is bigger than the difference (4.0 mL), so the difference could just be spread. **Question 7** (a) 50.8 g; (b) systematic error — every reading sits about 0.8 g high, the same push each time; (c) every repeat carries the same +0.8 g, so the mean of any number of repeats stays 0.8 g high; (d) zero the balance before use; calibrate it with known masses (or change the method). **Question 8** (a) systematic error; (b) every recorded temperature is 0.4 °C too high; (c) no effect — the +0.4 °C is in both readings of the pair, so it cancels in the fall; (d) much as it was for the *shape* of the curve (falls unchanged), though any claim about the actual temperatures (e.g. "cooled to 20.0 °C") would be 0.4 °C out. **Question 9** (a) A: dark soil warmed clearly more than light (31.2 vs 28.6 °C); B: almost no difference, and light slightly warmer (29.1 vs 29.8 °C); (b) any two — different soil moisture; different thermometer depth or placement; one thermometer uncalibrated; different time in the sun; (c) hold it — do not accept or reject it until the conflict is examined (compare methods, check calibration, gather more evidence). **Question 10** (a) any two of: same starting temperature each run; no draught; calibrated thermometer; layers wrapped the same way each time; (b) e.g. starting temperature — read it on the thermometer before every run and only begin at 80 °C; (c) if false, the results carry a shift that is not the layers' doing, so the conclusion about the layers rests on a faulty base (e.g. a draught on the 0-layer run would make the layers look better than they are). **Question 11** (a) the drop falls as layers are added, and the falls get smaller — a flattening curve; (b) 9.0, 5.5, 3.0 and 2.0 °C; (c) "more layers reduce the cooling" is supported (24.0 → 4.5 °C); "each extra layer saves the same amount" is not (9.0 vs 2.0 °C — the saving shrinks). **Question 12** (a) almost no gain from 2 to 4 hours (71 → 72) — "extra study makes little difference"; (b) scores rise steadily with study time (about 52 at 0.5 h to about 84 at 5 h) — a rising trend the two points hid; (c) bias in what evidence gets shown (cherry-picking); "is this all the evidence, or all the evidence that was shown?" **Question 13** (a) 10 g/L; (b) no plant was grown and nothing measured beyond 10 g/L, so 25 g/L is outside the data — any number there assumes the pattern continues; (c) the curve was already flattening (12.6 → 12.9 cm from 8 to 10 g/L), so the early steep rise the claim extends had already stopped. **Question 14** (a) a second factor differs (different soils), and watering was not controlled ("when remembered") — the sun/shade comparison is not fair; (b) the method is not written in full (no amounts, no rules), so another group cannot repeat it exactly; (c) one measurement per plant, no repeats — random error; (d) e.g. "In this single trial the sun plant grew taller, but with one plant per group and uncontrolled soil and water, the result does not yet support a general claim." **Question 15** (a) someone else follows the written method and gets results that agree; (b) any two — same drop heights and ball; same way of judging the top of the bounce; same number of repeats; instruments checked; (c) compare uncertainties and add repeats — if the two means sit within each other's uncertainty, the results actually agree. **Question 16** (a) random error — repeat more trials and use the mean, tighten the technique; (b) unrepresentative sample — choose sampling points at random, cover the whole creek; (c) systematic error — calibrate the instrument / check the zero; (d) not reproducible — write the method in full: amounts, times, rules for what counts.

Fully-worked solutions

Question 1 (a) The step immediately before EVALUATE is *Process and analyse* (Subtopics 4A, 4B, 4C). (b) The amber arrow leads back to *Plan* (Subtopics 3A, 3B, 3C). (c) Evaluation is placed before the conclusion is accepted because a conclusion deserves only as much trust as the method behind it — judging the method first finds the flaws a plan can still fix.

Question 2 (a) “Precise but not accurate”: readings 26.3–26.5 g, spread 0.2 g (precise), mean 26.38 g against the true 25.0 g (not accurate). (b) “Accurate but not precise”. The readings scatter widely (23.8–26.1 g) but they scatter *both ways* around the true value, so the high and low pushes cancel and the mean lands on 25.0 g. (c) Systematic error: the cluster is tight (little random scatter) yet the whole cluster sits about 1.4 g to one side — the signature of a steady push, not of chance scatter.

Question 3 (a) (1.5 + 1.7 + 1.55 + 1.65 + 1.6 + 1.7 + 1.5 + 1.6) ÷ 8 = 12.8 ÷ 8 = 1.60 s. (b) They sit both above and below the true 1.60 s — early on some trials, late on others. (c) Repeating and averaging. The right panel shows the running mean wobbling at first and settling onto 1.60 s as the eight readings are averaged — more repeats, tighter mean.

Question 4 (a) Stopwatch thumb → **random**: it is early on some trials and late on others, so the push changes direction. Balance not zeroed → **systematic**: it adds 0.3 g to *every* reading, the same push each time. (b) Random. A draught that blows “differently each time” changes direction and size from reading to reading — both-way scatter, not a steady shift. (c) Check the thermometer in melting ice (calibrate it) and apply the correction, or replace it with a checked one.

Question 5 (a) Check 1 (“was only the tested factor different?”) fails: hand stirring lets *stirring effort* differ between the beakers, and faster stirring speeds up dissolving. (b) Check 2 (“was the comparison fair?”) — one trial per temperature means a single, possibly unlucky pair of beakers. (c) Check 3 passes: the question asks about speed of dissolving, and the time taken to dissolve is exactly that quantity.

Question 6 (a) 42.0 − 38.0 = 4.0 mL. (b) With ± 2 mL the uncertainty bars (40–44 against 36–40 mL) just fail to overlap, and the difference is twice the uncertainty: against the spread of the data, 4.0 mL looks like a real difference, so the conclusion can stand. (c) With ± 6 mL the bars (36–48 against 32–44 mL) overlap heavily and the uncertainty is bigger than the difference: the 4.0 mL gap could simply be the spread of repeats, so the same conclusion is weak.

Question 7 (a) (50.8 + 50.7 + 50.9 + 50.8 + 50.8) ÷ 5 = 254.0 ÷ 5 = 50.8 g. (b) Systematic error. Every reading sits about 0.8 g above the true 50.0 g — the same push, same direction, every time, matching the 0.8 g empty-pan reading. (c) Because each repeat carries the same +0.8 g, the mean of twenty, fifty or a hundred repeats still carries it: repeating reduces random spread, not a steady shift. (d) Zero the balance before use, and/or calibrate it against known masses; if it cannot be zeroed, change the method (e.g. subtract the empty-pan reading from every measurement).

Question 8 (a) Systematic error — a fault in the instrument that pushes every reading the same way. (b) Each recorded temperature is 0.4 °C higher than the true temperature. (c) None. A fall is the difference of two readings, and both carry the same +0.4 °C, so it cancels: (24.4 − 20.4) = (24.0 − 20.0) = 4.0 °C. (d) For conclusions about the *cooling* (the falls, the shape of the curve) much as it was, because the falls are unchanged; for any conclusion stating an actual temperature (“cooled to 20.0 °C”) weaker, because every stated temperature is 0.4 °C out.

Question 9 (a) Group A found dark soil clearly warmer after an hour (31.2 °C against 28.6 °C). Group B found almost no difference, with light soil slightly warmer (29.1 against 29.8 °C). (b) Any two sensible ones: the soils differed in moisture between the groups; the thermometers sat at different depths or in different sun; one group’s thermometer was not calibrated; the hour in the sun differed (cloud passed). (c) Hold the conclusion. While conflicting evidence stands unexamined the conclusion is weak, so the class should compare methods, check calibration, and gather more evidence before accepting or rejecting “dark soil warms more”.

Question 10 (a) Any two of: the water started at the same temperature each run; no draught cooled one run more than another; the thermometer read truly; the layers were wrapped the same way each time. (b) Example for “same starting temperature”: read the thermometer before timing begins, on every run, and only start a run at 80 °C — the assumption then becomes a checked fact. (c) If the assumption is false, part of the difference between runs is not the layers’ doing. A draught on the zero-layer run, say, would inflate that run’s drop and make every layer look better than it truly is — the conclusion would rest on a shift that belongs to the fault, not to the insulation.

Question 11 (a) The drop falls as layers are added (24.0 → 4.5 °C), and the falls between successive layers get smaller — the curve flattens. (b) 24.0 − 15.0 = 9.0; 15.0 − 9.5 = 5.5; 9.5 − 6.5 = 3.0; 6.5 − 4.5 = 2.0 °C. (c) “More layers reduce the cooling” is strongly supported: the drop shrinks from 24.0 to 4.5 °C as layers go from 0 to 4. “Each extra layer saves the same amount” is *not* supported: the savings run 9.0, 5.5, 3.0, 2.0 °C — each added layer helps less than the one before.

Question 12 (a) Taken alone, (2, 71) and (4, 72) show almost no gain for two extra hours of study — "extra study time makes little difference". (b) Through all ten points the line of good fit rises steadily, with scores climbing from about 52 at 0.5 h to about 84 at 5 h — the full data shows a rising relationship the quoted pair sat inside and hid. (c) Bias in the choice of what evidence is shown (cherry-picking). The standing question: *is this all the evidence, or all the evidence that was shown?*

Question 13 (a) 10 g/L. (b) Nothing was measured between 10 and 25 g/L — no plant was grown there — so the 25 g/L figure sits in a range with no data under it; any height quoted there assumes, without evidence, that the pattern continues. (c) Inside the measured range the curve was flattening: growth rose 12.0 → 12.6 → 12.9 cm at 6, 8 and 10 g/L. A pattern already flattening gives no support to extending the early steep straight line, so even the shape of the data argues against the 55 cm claim.

Question 14 (a) The two plants differed in a second factor (different soils), and watering was left uncontrolled ("when remembered") — so the sun/shade comparison is not a fair test and the method is not valid. (b) The method states no amounts, no soil volumes, no watering rule — another group cannot follow it exactly, so the investigation is not reproducible. (c) One height measurement per plant and no repeat trials: random error (and a sample of one) goes unaddressed; repeats and more plants would reduce it. (d) A conclusion its evidence can carry: "In this one trial the sun plant grew taller (31 cm against 24 cm), but with one plant per group, different soils and uncontrolled watering, this does not yet show that light caused the difference."

Question 15 (a) An investigation is reproducible when someone else follows the written method and gets results that agree. (b) Any two: were the same drop heights and the same ball used; was the bounce judged the same way (top of the bounce, eye level); were the same number of repeats taken and averaged; were the rulers checked against each other. (c) Add repeats and compare uncertainties: if each group's mean sits inside the other's ± uncertainty, the results agree after all; if not, hunt for a remaining method or instrument difference.

Question 16 (a) Random error → repeat more trials and use the mean; tighten the measuring technique. (b) Unrepresentative sample → choose sampling points at random; cover the whole area or group (the whole creek, not the footbridge spot). (c) Systematic error → calibrate the instrument / check the zero (melting-ice check) or change the method. (d) Not reproducible → write the method in full: amounts, times, rules for what counts ("heat for 5 minutes", not "heat for a while").

Where this leads: the conclusion and report steps of the inquiry sequence (Subtopics 5B, 6A and 6B) take over from here — a conclusion is only as good as the evaluation behind it — and the insulation story returns later in the book, where the *why* of insulating layers (heat transfer) is finally taught. Everything evaluated in this subtopic — validity, reproducibility, error, bias, assumptions, conflicting evidence and areas of uncertainty — is also what the Chapter 5 test and the investigation tasks ask you to apply.

Building an evidence-based argument, assessing claims, and the ethics and cultural protocols of using secondary sources

Chapter 5 — Evaluating (Science, Levels 9 and 10)

Subtopic 5B · VC2 Science Inquiry Skills — Evaluating

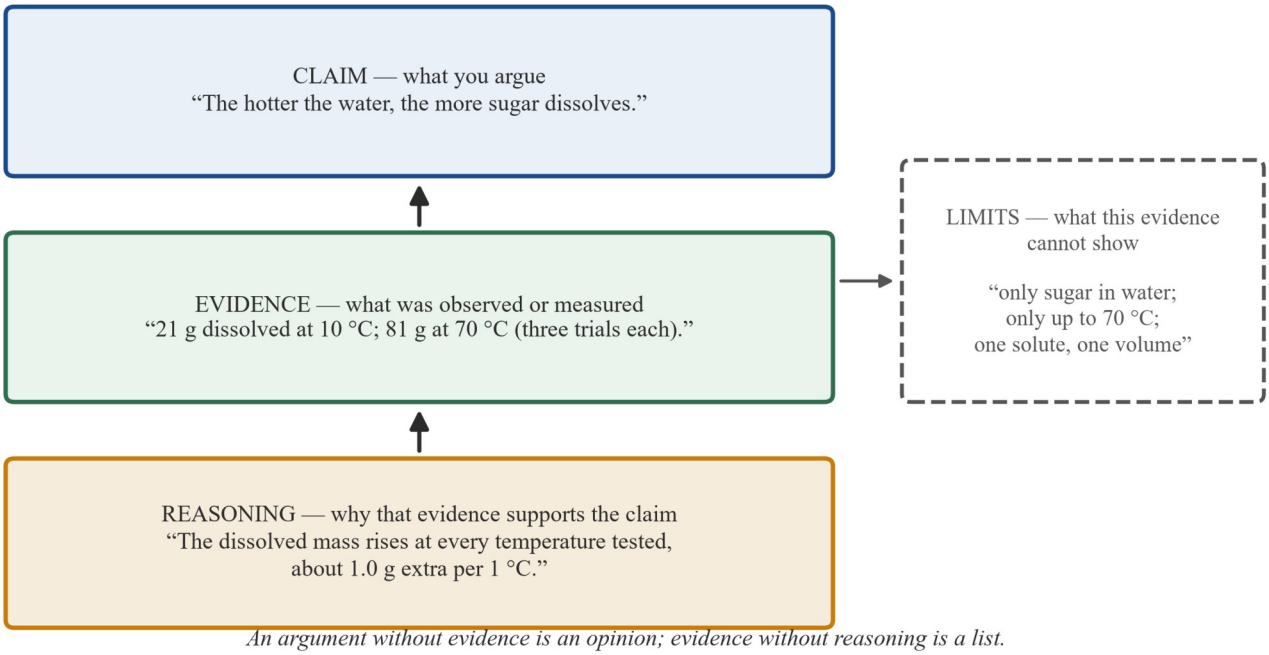
CD code: `VC2S10I07` (delivered whole by this subtopic) **Content description (verbatim):** "arguments based on a variety of evidence can be constructed to support conclusions or evaluate claims, including consideration of any ethical issues and cultural protocols associated with accessing, using or citing secondary data or information" **Elaboration ids drawn on:** `VC2S10I07.E01` · `VC2S10I07.E02` · `VC2S10I07.E03` · `VC2S10I07.E04` · `VC2S10I07.E05` **Achievement standard (verbatim):** "They provide evidence-based explanations for findings and construct logical arguments based on the evaluation of multiple sources of evidence to justify conclusions and assess claims." **Maths interlocks (D11):** `VC2M10ST04` — analyse claims, inferences and conclusions of statistical reports in the media and other places, by linking claims to displays, statistics and representative data, including ethical considerations and identification of potential sources of bias (the "reading a statistic" checks in this subtopic are exactly that skill).

All datasets in this subtopic are simulated data for practice.

Theory

Subtopic 5A judged investigations. This one builds and assesses arguments. Subtopic 5A looked back over an investigation and asked whether its method was valid, its data reproducible, and its conclusion as strong as the evidence allowed. This subtopic takes the next step: once the evidence is in, what do you *say* with it? You will construct **evidence-based arguments** (arguments where a claim is held up by evidence and reasoning), assess claims made by other people — in the media, in advertisements, in other groups' reports — and use **secondary sources** (sources of data or information you did not collect yourself) honestly and lawfully. The two halves belong together: the same care that goes into building your own argument is the care you use to test someone else's.

The parts of an evidence-based argument. A scientific argument has three working parts and a fourth that keeps it honest. The **claim** is what you argue — the statement you want your reader to accept. The **evidence** is what was observed or measured — the data, quoted with its numbers and units. The **reasoning** is the part that does the real work: it says *why* that evidence supports that claim, so the reader can follow the thinking rather than being asked to take your word for it. And the **limits** state what the evidence cannot show — which conditions, substances or ranges the argument does not reach. Stating limits is not weakness; it is what makes an argument trustworthy.

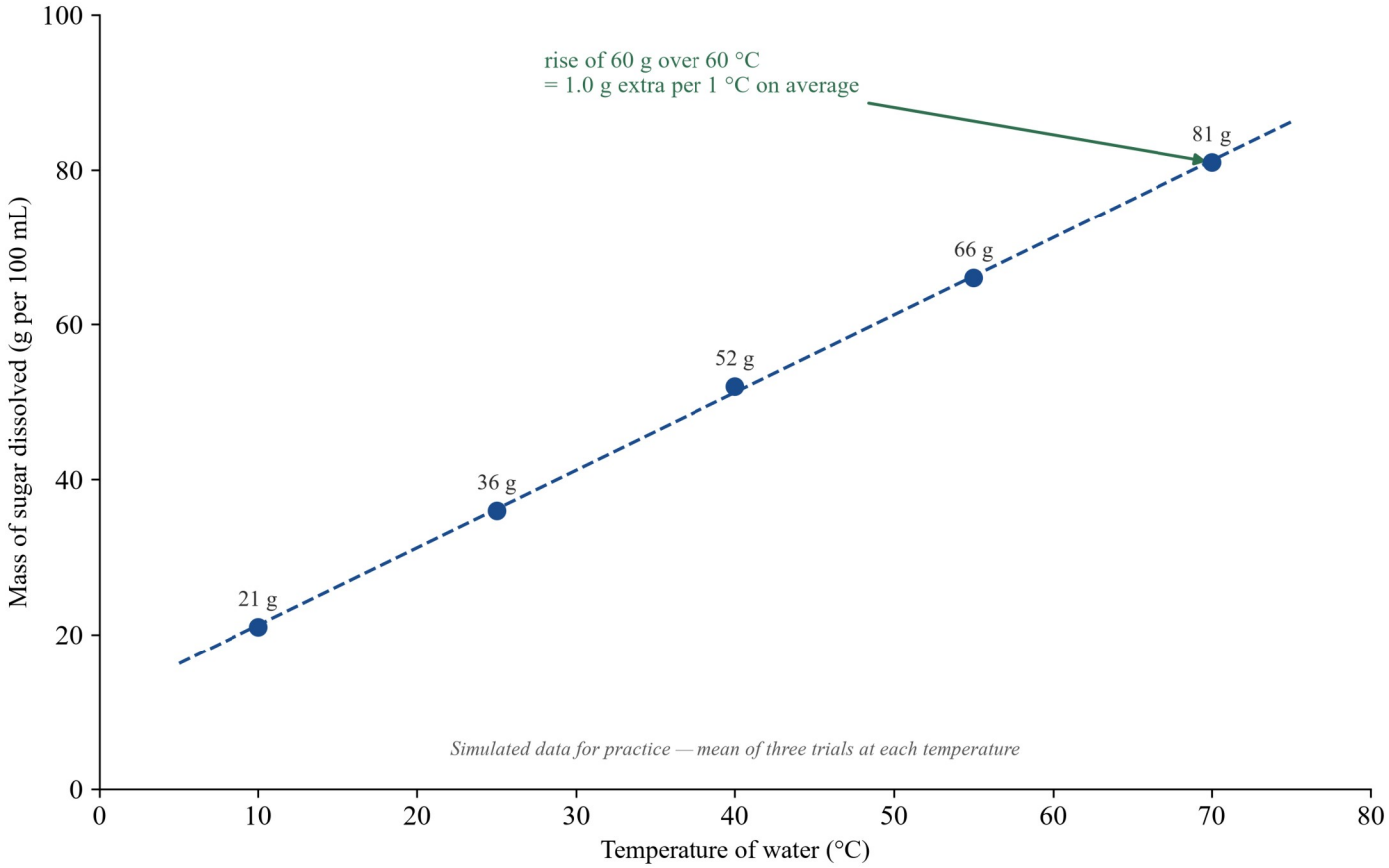


An argument without evidence is an opinion; evidence without reasoning is a list.

The parts of an evidence-based argument as three stacked boxes joined by upward arrows: the blue CLAIM box reads “The hotter the water, the more sugar dissolves”; the green EVIDENCE box reads “21 g dissolved at 10 °C; 81 g at 70 °C (three trials each)”; the amber REASONING box reads “The dissolved mass rises at every temperature tested, about 1.0 g extra per 1 °C”; a dashed grey box to the right labelled LIMITS reads “only sugar in water; only up to 70 °C; one solute, one volume”; caption: an argument without evidence is an opinion; evidence without reasoning is a list

Read the figure top to bottom. The claim says *what*; the evidence says *what was found*; the reasoning connects the two — notice it quotes the pattern in the numbers (rising at every temperature, about 1.0 g per 1 °C), not just “the results show it”. The dashed limits box keeps the claim inside the evidence: nothing here was measured for any substance but sugar, or above 70 °C, so the claim must not talk about anything else. That is the whole shape. Every argument you write in this book — and every claim you assess — can be sorted into these four parts.

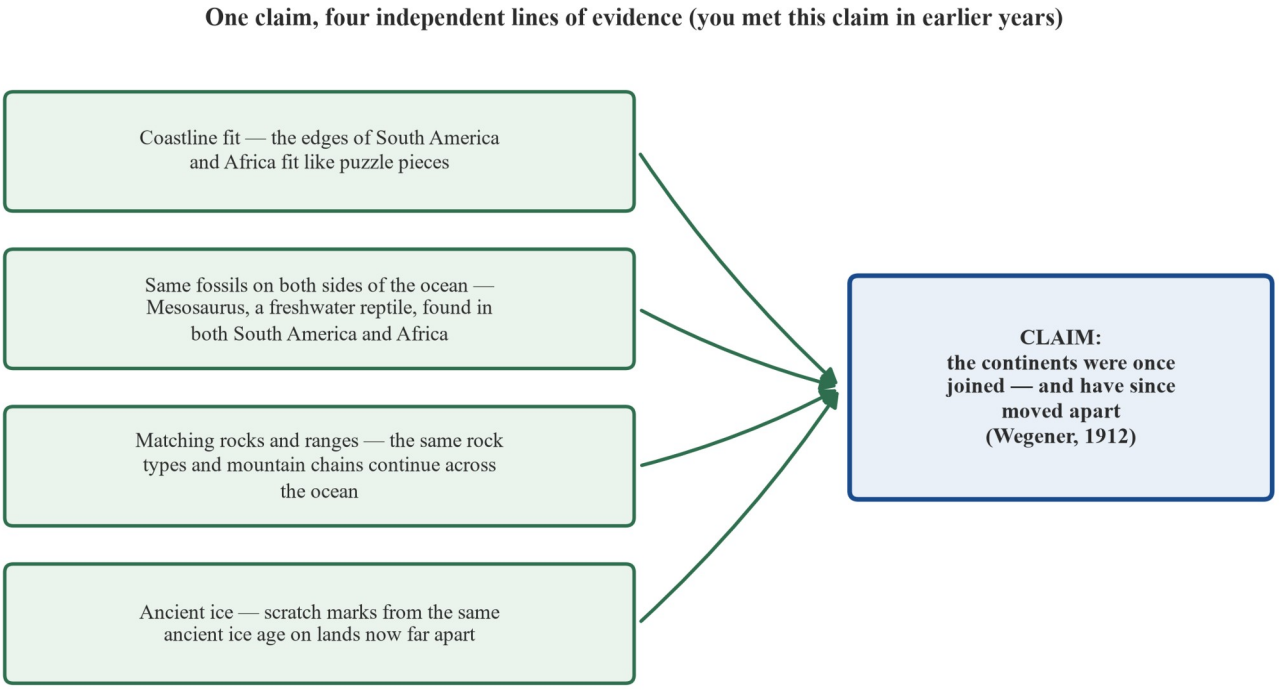
Justifying a conclusion from your own data (elaboration `VC2S10I07.E04`). The commonest argument you will build is the one at the end of your own investigation: *what conclusion does this data support?* A class investigated the question *does hotter water dissolve more sugar?*, measuring the mass of sugar that dissolved in 100 mL of water at five temperatures, with three trials at each temperature. Their primary data — data the class collected itself — is plotted below.



A scatter plot of simulated class data for practice: mass of sugar dissolved in 100 mL of water against temperature, with points at 10 °C = 21 g, 25 °C = 36 g, 40 °C = 52 g, 55 °C = 66 g and 70 °C = 81 g, a dashed blue line of best fit rising steadily through them, and a green note "rise of 60 g over 60 °C = 1.0 g extra per 1 °C on average"; axis labels: Temperature of water (°C) and Mass of sugar dissolved (g per 100 mL)

The data rises at every temperature tested, so the conclusion it can carry is: *the hotter the water, the more sugar dissolves*. A good conclusion also says *how much*: from 10 to 70 °C the dissolved mass rose 81 − 21 = 60 g over a rise of 60 °C, an average of 1.0 g extra per 1 °C. Notice what justifying a conclusion is *not*: it is not choosing the most interesting sentence you can think of and hoping. It is pointing at features of the data — the direction of the trend, the size of the rise — and showing that the conclusion fits inside what was actually measured. Subtopic 5A's rule still binds: the conclusion only deserves as much trust as the method behind it, so the evaluation comes first, the conclusion second.

A range of evidence behind one claim (elaboration `VC2S10I07.E03`). Some claims rest on many separate lines of evidence rather than one dataset. You met the classic example in earlier years: the claim that the continents were once joined. In 1912, Alfred Wegener argued for it using several lines of evidence that had nothing to do with each other except where they pointed.



Any one line alone is only a hint. Four independent lines pointing the same way is an argument.

Four green evidence cards on the left — coastline fit (South America and Africa fit like puzzle pieces); the same freshwater reptile fossil Mesosaurus found in both South America and Africa; matching rock types and mountain chains continuing across the ocean; scratch marks from the same ancient ice age on lands now far apart — each with an arrow converging on a blue CLAIM box reading "the continents were once joined — and have since moved apart (Wegener, 1912)"; caption: any one line alone is only a hint; four independent lines pointing the same way is an argument

No one line proves the claim by itself. A coastline shape is suggestive, but coastlines change. A fossil found on two continents is strange, but there might be another explanation — land bridges, drifting seeds of explanation. Matching rocks could be coincidence. What makes it an argument is that four *independent* lines — shape, fossils, rocks, ancient ice — all point the same way, and the joined-continents claim explains all four at once. That is the pattern elaboration `VC2S10I07.E03` is after, whatever the claim is about: a range of evidence, converging. Written out in full, the argument looks like this.

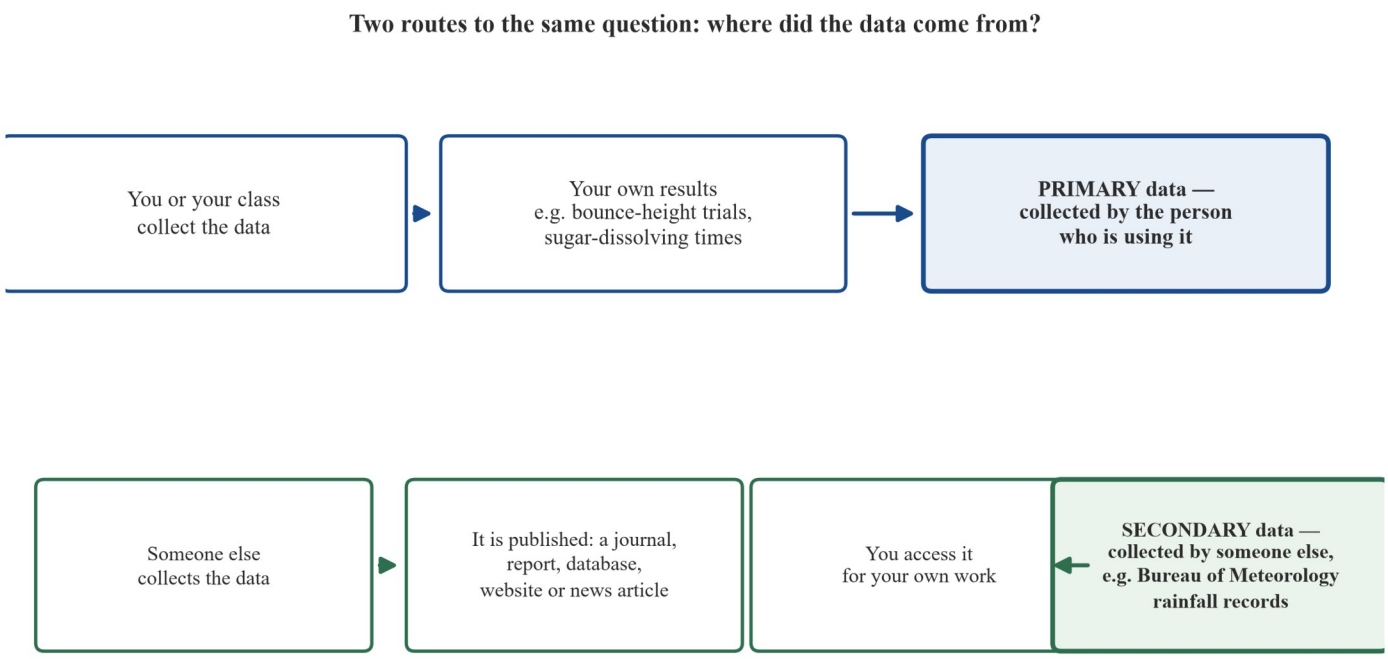
A complete evidence-based argument, written out

CLAIM	The continents were once joined, and have since moved apart.
EVIDENCE 1	The coastlines of South America and Africa fit together like puzzle pieces.
EVIDENCE 2	The same freshwater reptile fossil, Mesosaurus, is found in both South America and Africa — it could not have crossed an ocean.
EVIDENCE 3	The same rock types and mountain chains continue across the ocean.
REASONING	Three independent lines of evidence point the same way. No one line proves the claim alone — a coastline shape is only a hint, and a shared fossil could have another cause — but together they make the joined-continents claim the best explanation.
LIMITS	Wegener could not yet say how continents move; that came later. The argument stood on its evidence even without a mechanism.

A complete written argument in labelled rows: CLAIM "The continents were once joined, and have since moved apart"; EVIDENCE 1 "The coastlines of South America and Africa fit together like puzzle pieces"; EVIDENCE 2 "The same freshwater reptile fossil, Mesosaurus, is found in both South America and Africa — it could not have crossed an ocean"; EVIDENCE 3 "The same rock types and mountain chains continue across the ocean"; REASONING "Three independent lines of evidence point the same way; no one line proves the claim alone, but together they make the joined-continents claim the best explanation"; LIMITS "Wegener could not yet say how continents move; that came later — the argument stood on its evidence even without the explanation"

Two things are worth copying from this model. First, the reasoning row does not just repeat the evidence — it says what each line does and does not do, and why together they are enough. Second, the limits row admits what was missing: Wegener could not yet say *how* continents move, and the idea waited decades for the explanation to catch up with the evidence. An argument can be strong on evidence even before the full explanation exists — but you must say so, in the limits.

Primary and secondary data — where did the data come from? You have just seen an argument built from data the class collected itself. That is **primary data**: data collected by the person (or group) who is using it. Much of science runs on the other kind. **Secondary data** is data collected by someone else that you access for your own work — published results, databases, records kept over many years, museum collections, journal articles, or a news article reporting any of these.

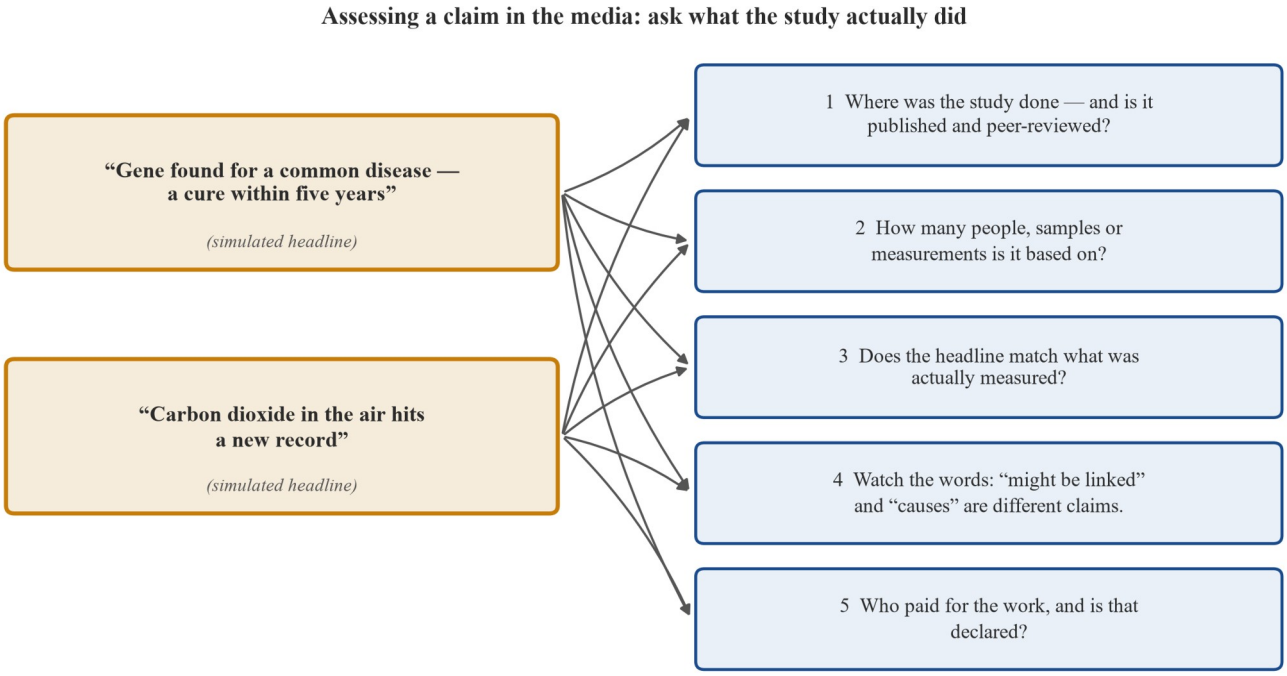


The same numbers are primary for the people who collected them and secondary for everyone who reads them later.

Two routes to the same question drawn as horizontal lanes: the top lane, "You or your class collect the data" leading to "Your own results (e.g. bounce-height trials, sugar-dissolving times)" leading to a blue box "PRIMARY data — collected by the person who is using it"; the bottom lane, "Someone else collects the data" leading to "It is published: a journal, report, database, website or news article" leading to "You access it for your own work" leading to a green box "SECONDARY data — collected by someone else, e.g. Bureau of Meteorology rainfall records"; caption: the same numbers are primary for the people who collected them and secondary for everyone who reads them later

The distinction is about *who collected it*, not about what it is. The Bureau of Meteorology's rainfall records are primary data for the Bureau — and secondary data for you the day you download them for a project. Secondary sources are not second-rate: they let you use measurements you could never make yourself (a hundred years of rainfall, air measured since 1976). But because you did not collect the data, you inherit two questions you never have to ask of your own data: *can I trust this source?* and *am I allowed to use it this way?* The rest of the theory answers both.

Assessing claims in the media (elaborations `VC2S10I07.E01` and `VC2S10I07.E02`). Most claims you meet in your life will arrive as headlines, not as lab reports. The skill is not to refuse to believe anything; it is to ask what the study behind the headline actually did. Five questions cover most of it.



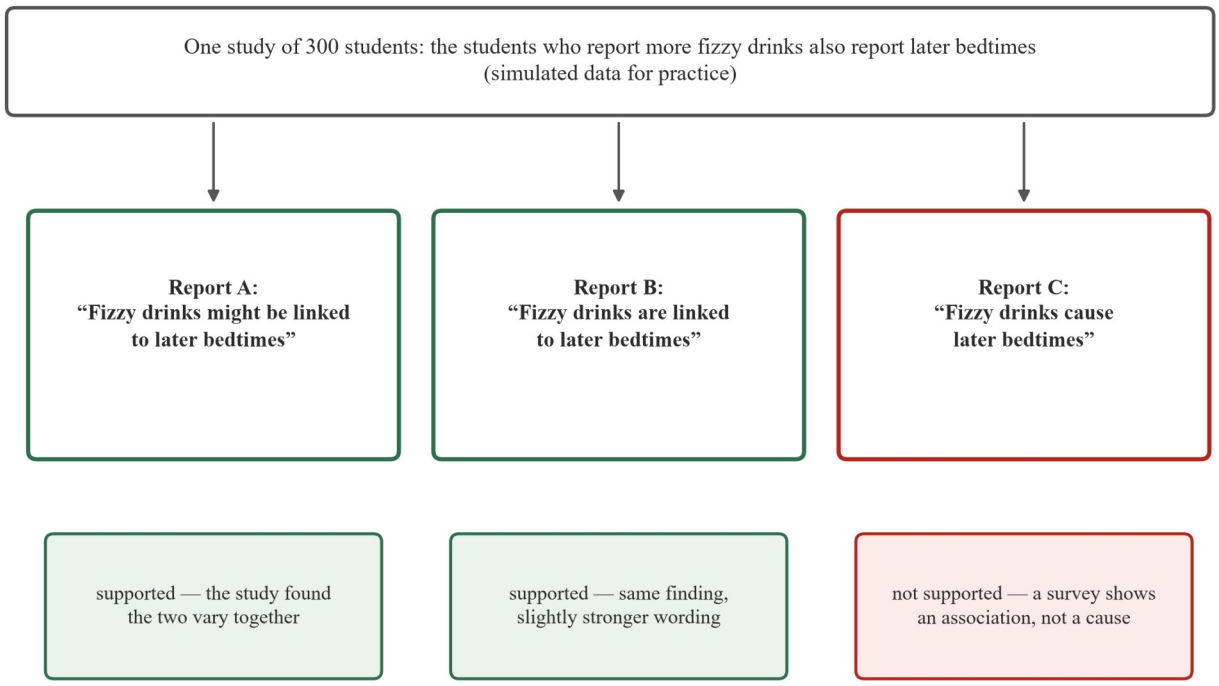
Two amber headline cards on the left — "Gene found for a common disease — a cure within five years" and "Carbon dioxide in the air hits a new record", both labelled "simulated headline" — with grey arrows from each pointing to five blue check boxes on the right: 1 Where was the study done — and is it published and peer-reviewed? 2 How many people, samples or measurements is it based on? 3 Does the headline match what was actually measured? 4 Watch the words: "might be linked" and "causes" are different claims. 5 Who paid for the work, and is that declared?

Run the checks on the first (simulated) headline. *Gene found* is a finding; *cure within five years* is a prediction nobody has made in a lab — check 3 catches it, because the headline sells something the study did not measure. Check 2 asks how big the study was; a gene study of a few families is not the same claim as one of tens of thousands of people. Check 5 matters because a study paid for by a company that sells the treatment deserves extra scrutiny — the funding does not make the study wrong, but you are entitled to know.

Now the second headline: *Carbon dioxide in the air hits a new record.* Where would you even begin to check that? Elaboration `VC2S10I07.E02` gives the move: research the *method* behind the number. In Australia, one of the longest direct records of carbon dioxide in the air comes from the Cape Grim Baseline Air Pollution Station in north-west Tasmania, run by the Bureau of Meteorology and CSIRO, which has measured carbon dioxide in clean ocean air continuously since 1976. So a reader can ask: is the "record" from a real, long-running record like that — or from a few measurements in one city street? And one more thing, check 3 again: a record at one station is a fact about *that station*; a claim about the whole atmosphere needs measurements from many places around the world. The headline is only as good as the method standing behind its number.

CSIRO and Bureau of Meteorology, "Cape Grim Baseline Air Pollution Station", 2026. csiro.au — retrieved 21 August 2026.

Might be linked, is linked, causes. The same finding can be reported three ways, and only some of them are fair. Suppose one study of 300 students found that the students who reported more fizzy drinks also reported later bedtimes (simulated data for practice). Watch what happens to the wording.



Association: two things vary together (Subtopic 4C). Cause: one makes the other happen — that needs a fair test, not a survey.

One finding reported three ways: a top box states "One study of 300 students: the students who report more fizzy drinks also report later bedtimes (simulated data for practice)"; below it three cards — Report A "Fizzy drinks might be linked to later bedtimes" with green verdict "supported — the study found the two vary together"; Report B "Fizzy drinks are linked to later bedtimes" with green verdict "supported — same finding, slightly stronger wording"; Report C "Fizzy drinks cause later bedtimes" with red verdict "not supported — a survey shows an association, not a cause"; caption: association means two things vary together; cause means one makes the other happen — that needs a fair test, not a survey

Reports A and B stay inside the finding: the two things vary together, and saying so is fair. Report C steps outside it. The study was a survey — it asked students what they did; it tested nothing. Students who drink more fizzy drinks might also differ in a dozen other ways (screen time, sport, age), and any of those could sit behind the later bedtimes. A survey can show an **association** — two things varying together, the idea you met in Subtopic 4C — but it cannot show **cause**. For cause, you need a fair test (Subtopic 3A's conditions: only the tested factor differs). Elaboration `VC2S10I07.E01` asks exactly this of media reports: what did the study actually do, and how might the public read the wording? The difference between "might be linked" and "causes" is small on the page and enormous in what people believe.

Trusting a secondary source: six questions. Before secondary data goes into your work, run the source itself through a check. The six questions below are the habit to build — they are the same questions, in a different order, that Subtopic 5A aimed at methods.

Six questions to ask before you trust a secondary source

- ☐ 1 Who made it — and why are they a good source on this?
- ☐ 2 Does it show its evidence, so you can check it?
- ☐ 3 How old is it? Would newer data change the claim?
- ☐ 4 Do other sources, working separately, agree with it?
- ☐ 5 Is it declared who paid for the work?
- ☐ 6 Are there ethical or cultural permissions needed to use it?

A "no" does not always kill a source — but you must know the answer before you cite it.

A blue card titled "Six questions to ask before you trust a secondary source", listing with tick boxes: 1 Who made it — and why are they a good source on this? 2 Does it show its evidence, so you can check it? 3 How old is it? Would newer data change the claim? 4 Do other sources, working separately, agree with it? 5 Is it declared who paid for the work? 6 Are there ethical or cultural permissions needed to use it?; caption: a "no" does not always kill a source — but you must know the answer before you cite it

A few of the six deserve a second look. Question 1 is about fit, not fame: a museum is a good source on natural history, a weather bureau on weather — the same source can be strong on one question and weak on another. Question 3 matters because science moves: a claim about the air, the climate or the sky from 1990 may simply be out of date. Question 4 is the strongest check of all — when sources that have nothing to do with each other agree, the claim is much harder to explain away. And question 6 is the one most people forget; it gets its own treatment at the end of this theory.

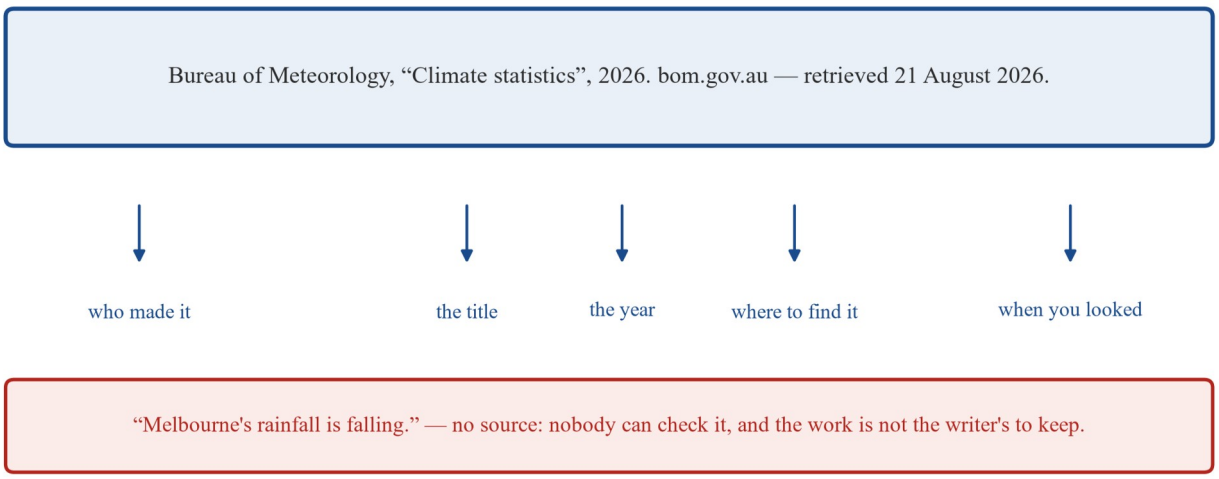
Reading the statistic behind the claim (Maths interlock `VC2M10ST04`). Many claims lean on a single statistic, and the statistic is only as good as the sample behind it. Consider the claim *"9 out of 10 students prefer Friday practicals."*

Two panels comparing the statistic with the sample behind it: left panel, amber claim box “9 out of 10 students prefer Friday practicals”, below it “Who was actually asked? 10 students — all from one class”, and a red verdict “a small sample from one group — 9 out of 10 here tells you about ten people, not about the school”; right panel, amber box “Whole year level asked: 312 of 540 prefer Friday practicals”, below it “312 ÷ 540 = 58% — most, but not 9 out of 10 (90%)”, and a green verdict “a large sample of the whole group — ‘most students prefer Fridays’ is fair; ‘9 out of 10’ is not”

Nine out of ten sounds like nearly everyone — it is 90%. But look at who was asked: ten students from one class. Ten people cannot speak for a school, and one class might love Fridays for reasons that have nothing to do with practicals (a favourite teacher, a timetable quirk). When the whole year level was asked, 312 out of 540 said the same — which is 312 ÷ 540, about 58%: most students, genuinely, but nowhere near 9 out of 10. The fair claim from the big poll is *“most students prefer Friday practicals”*; the “9 out of 10” version belongs to ten people only. Whenever a claim hands you a statistic, ask the four questions of `VC2M10ST04`: *how many were measured? who, exactly? how were they chosen? and how was the question asked?* A statistic from a small, hand-picked, or one-corner sample tells you about the sample — not about the group it claims to describe.

Using secondary sources ethically (elaboration `VC2S10I07.E05`). Once you decide a source is trustworthy, three obligations come with using it. They are not good manners; they are part of doing science.

One — acknowledge it. Every piece of someone else’s data, idea or words that enters your work must carry an acknowledgement — a **citation**, a line that tells the reader exactly what you used and where it came from. Using someone else’s work as if it were your own is called **plagiarism**, and it is the one writing failure that cannot be argued away. The house format for a source line has five parts, and each has a job.

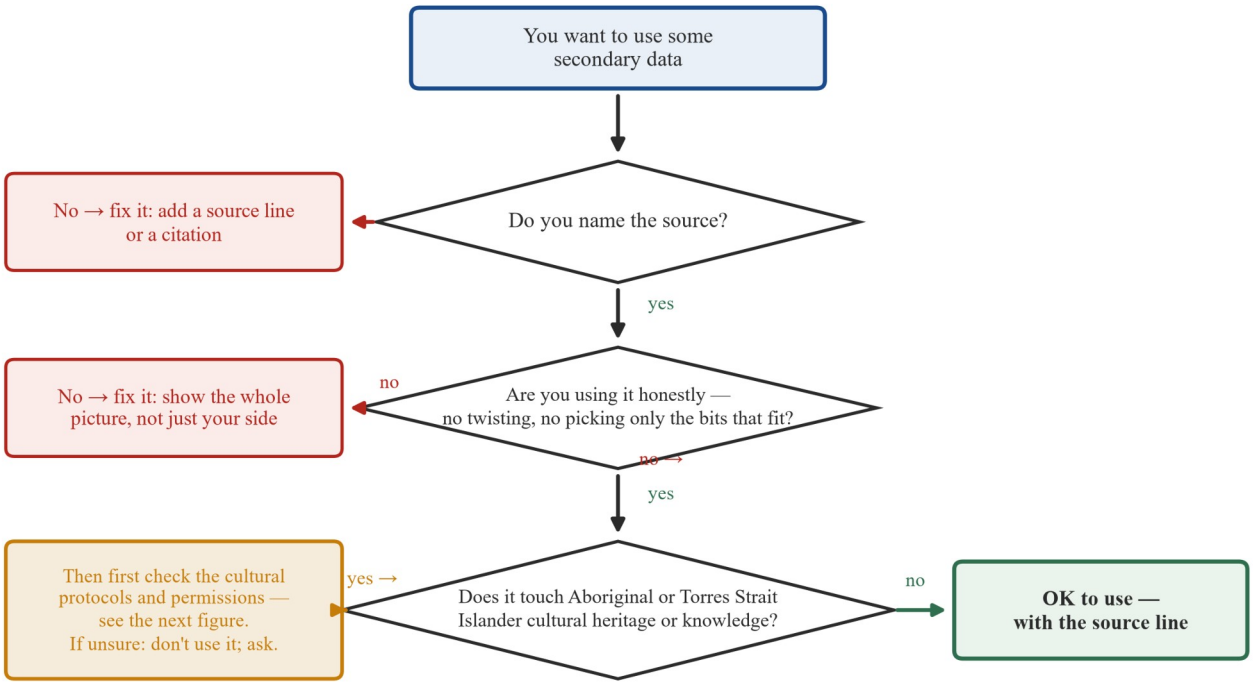


A labelled source line: “Bureau of Meteorology, ‘Climate statistics’, 2026. bom.gov.au — retrieved 21 August 2026” with arrows pointing down to five labels — who made it; the title; the year; where to find it; when you looked; below, a red box reads “‘Melbourne’s rainfall is falling.’ — no source: nobody can check it, and the work is not the writer’s to keep”

The retrieval date earns its place: a web page can change after you read it, and your reader deserves to know which version you saw. A claim with no source line fails two of the six source questions at once — you cannot say who made it, and nobody can check it.

Two — use it honestly. Citing a source and then twisting what it says is its own kind of dishonesty. That includes quoting only the results that suit your claim while hiding the ones that do not — the cherry-picking you judged in Subtopic 5A — and reporting a source’s “might be linked” as a “causes”. The rule is simple: show the whole picture, including the parts that work against you.

Three — check permissions. Some data comes with conditions on its use: copyright on images and text, terms of use on databases, and — for material touching Aboriginal and Torres Strait Islander heritage and knowledge — cultural protocols. The flow below runs all three obligations in order every time you reach for secondary data.

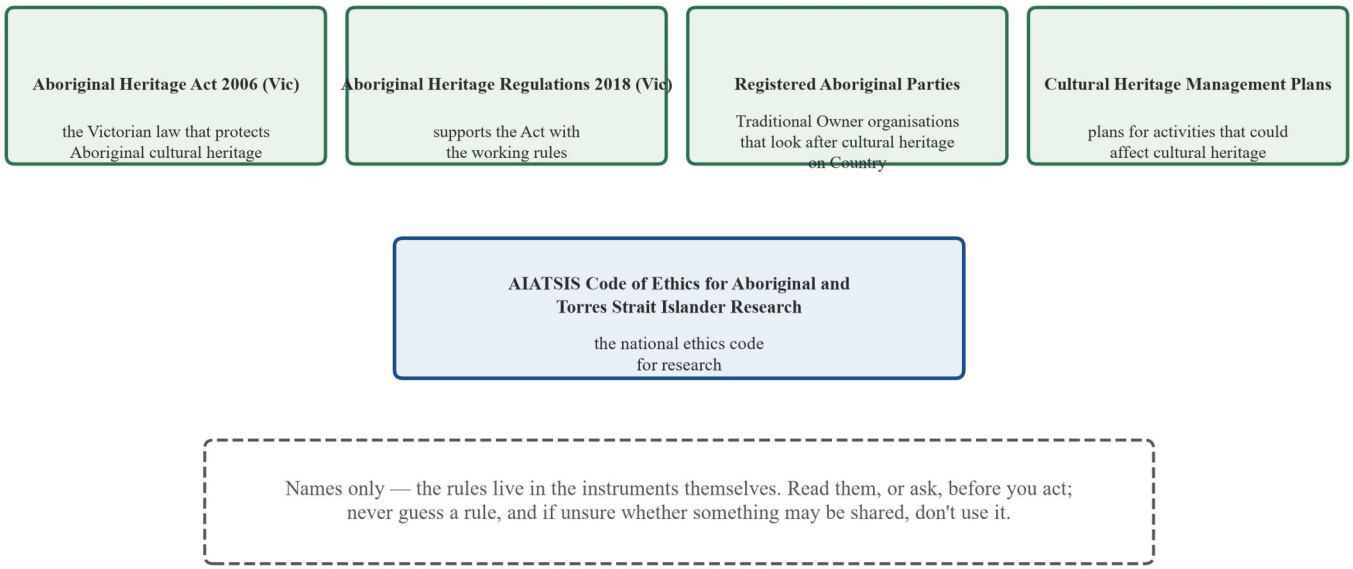


The same three questions every time — before the data goes into your work.

A decision flowchart: top box "You want to use some secondary data" leading to diamond "Do you name the source?" — no, red box "fix it: add a source line or a citation"; yes, to diamond "Are you using it honestly — no twisting, no picking only the bits that fit?" — no, red box "fix it: show the whole picture, not just your side"; yes, to diamond "Does it touch Aboriginal or Torres Strait Islander cultural heritage or knowledge?" — yes, amber box "Then first check the cultural protocols and permissions — see the next figure. If unsure: don't use it; ask"; no, green box "OK to use — with the source line"; caption: the same three questions every time, before the data goes into your work

Cultural protocols — the second obligation. VCAA's elaboration `VC2S10I07.E05` defines cultural protocols as *"ethical principles that guide behaviour in particular situations when interacting with Aboriginal and/or Torres Strait Islander Peoples and that are designed to protect Aboriginal and Torres Strait Islander cultural and intellectual property rights."* Read that definition twice: cultural protocols are about *protection* — they exist so that Aboriginal and Torres Strait Islander cultural and intellectual property stays under the control of the people it belongs to. Some Aboriginal and Torres Strait Islander knowledge is not for everyone to share, and whether something may be shared, and with whom, is decided by its owners — not by the person who found it on a website.

For science work in Victoria, there is a framework of published instruments you should know by name. The **Aboriginal Heritage Act 2006 (Vic)** is the Victorian law that protects Aboriginal cultural heritage, and the **Aboriginal Heritage Regulations 2018 (Vic)** supports it with the working rules. Under this framework, **Registered Aboriginal Parties** are Traditional Owner organisations that look after cultural heritage on Country, and **Cultural Heritage Management Plans** are prepared for activities that could affect cultural heritage. For research across Australia, the **AIATSIS Code of Ethics for Aboriginal and Torres Strait Islander Research** sets the ethical standards.



Five green cards naming the instruments — Aboriginal Heritage Act 2006 (Vic), the Victorian law that protects Aboriginal cultural heritage; Aboriginal Heritage Regulations 2018 (Vic), which supports the Act with the working rules; Registered Aboriginal Parties, Traditional Owner organisations that look after cultural heritage on Country; Cultural Heritage Management Plans, plans for activities that could affect cultural heritage — and one blue card, the AIATSIS Code of Ethics for Aboriginal and Torres Strait Islander Research, the national ethics code for research; caption: names only — the rules live in the documents themselves; read them, or ask, before you act

Parliament of Victoria, "Aboriginal Heritage Act 2006 (Vic)". legislation.vic.gov.au — retrieved 21 August 2026. Parliament of Victoria, "Aboriginal Heritage Regulations 2018 (Vic)". legislation.vic.gov.au — retrieved 21 August 2026. AIATSIS, "AIATSIS Code of Ethics for Aboriginal and Torres Strait Islander Research", 2020. aiatsis.gov.au — retrieved 21 August 2026.

What does this mean for your work? Three habits, no exceptions. First, if secondary material touches Aboriginal or Torres Strait Islander heritage or knowledge — an image, a story, data about Country or heritage places — treat the cultural protocols as part of the permissions check, not an afterthought. Second, know the names above, and read the documents before you act on them: the rules live in the instruments themselves, never in someone's memory of them, and you must never guess at a rule. Third, hold the safest rule in science when unsure: if you do not know whether something may be shared or used, *don't use it — ask.* Your teacher, and the relevant published instrument, are where the answer comes from.

Activities

Activity 1 — Build the drift argument. Working from the evidence-chain figure and without looking at the written-argument figure, work in pairs to write out the continental-drift argument in full: claim, one row per line of evidence, a reasoning row that says what each line does and does not show, and a limits row. Then compare your version with the written-argument figure. Mark one place where your reasoning said more than the model's, one place where it said less, and fix the limits row together.

Activity 2 — The headline clinic (elaborations `VC2S10I07.E01` and `VC2S10I07.E02`). Take the two simulated headlines of the headline-check figure (or two real headlines your teacher provides). For each: (1) run all five checks and record what each check finds; (2) write two questions you would need answered before you could trust the claim; (3) rewrite the headline so it says only what the study behind it actually showed — wording included

("might be linked" stays "might be linked"). Swap headlines with another pair and run their rewritten version back through check 3.

Activity 3 — Source line and verdict (elaboration `VC2S10I07.E05`). Choose three secondary sources your class is allowed to open (school library pages, the Bureau of Meteorology, Museums Victoria). For each: (1) write a full source line in the five-part format — who made it, title, year, where to find it, when you looked; (2) run the six source questions and record each answer; (3) give a verdict — *use*, *use with care* (state why), or *don't use until you ask*. As a class, collect the verdicts on the board: notice how often question 4 (do other sources agree?) is the one that cannot be answered from a single page.

Worked examples

Datasets in this section are simulated data for practice.

Example 1 — scaffolded

A class measures the mass of sugar that dissolves in 100 mL of water at five temperatures (three trials each, means shown): 21 g at 10 °C, 36 g at 25 °C, 52 g at 40 °C, 66 g at 55 °C, 81 g at 70 °C.

(a) State the trend in the data.
(b) Calculate the average rise in dissolved mass per 1 °C across the full range.
(c) Which conclusion can the data carry — *"the hotter the water, the more sugar dissolves"* or *"hotter water dissolves more of any solid"*? Justify your choice.

Working	Comment
(a) The dissolved mass rises at every temperature tested: 21 → 36 → 52 → 66 → 81 g	The trend is stated from the numbers, not from a feeling.
(b) Rise in mass: 81 − 21 = 60 g; rise in temperature: 70 − 10 = 60 °C; 60 ÷ 60 = 1.0 g per 1 °C	A conclusion is stronger when it says *how much*, not just *which way*.
(c) The first conclusion. The data was measured for sugar only, at these five temperatures, up to 70 °C	The second conclusion reaches beyond the data — no other solid was tested, so the data cannot speak for "any solid".

The data supports *"the hotter the water, the more sugar dissolves"*, rising about 1.0 g per 1 °C on average. The conclusion fits inside what was measured; the broader claim does not, because nothing but sugar was ever tested.

Example 2 — scaffolded

A student council report says: *"9 out of 10 students prefer Friday practicals — the Friday practical program should continue."* The 9 out of 10 came from asking 10 students in one class. A later poll of the whole year level found 312 of 540 students preferred Friday practicals.

(a) Write 9 out of 10 and 312 out of 540 as percentages.
(b) Explain why the 9 out of 10 figure does not describe the year level.
(c) Write a claim the whole-year poll can actually carry.

Working	Comment
(a) 9 ÷ 10 = 90%; 312 ÷ 540 ≈ 58%	Put both claims on the same scale before comparing them.
(b) The 9 out of 10 came from 10 students in one class — a small sample from one corner of the year level, who may share reasons nothing else in the year level shares	Sample size and *who was asked* decide how far a statistic reaches.
(c) "Most students prefer Friday practicals" (about 58% of the year level)	The fair claim matches the size of the evidence behind it.

The 90% figure is real but it describes ten people; the whole-year poll supports *most* (58%), not *9 out of 10*. The statistic did not change the truth — the sample did.

Example 3 — unscaffolded

Assess the simulated headline *"Carbon dioxide in the air hits a new record"* the way elaboration `VC2S10I07.E02` directs: by researching the method behind the number. What questions would you ask, and what could a real long-running record such as the Cape Grim Baseline Air Pollution Station support — and not support?

Start with check 2 and check 3: *where and how was carbon dioxide measured, and for how long?* A "record" claim is only checkable against a long, steady record — and that is exactly what Cape Grim provides: the station in north-west Tasmania, run by the Bureau of Meteorology and CSIRO, has measured carbon dioxide directly in clean ocean air continuously since 1976, so a new highest reading there is a genuine fact about that record, not a one-off guess. But the method also sets the limit of the claim: Cape Grim measures clean ocean air at one place, so its record supports a statement about that station's measurements — a claim about the whole atmosphere needs the agreement of stations around the world. So the fair assessment is: the headline *could* be fair, if the record it quotes is a long direct record like Cape Grim's and if "the air" is backed by measurements from many places; until both are checked, the headline is a claim waiting for its method.

Example 4 — unscaffolded

Construct an evidence-based argument from the following primary data (simulated data for practice): a ball is dropped from five heights and the bounce height measured, three trials each, means shown — drop 20 cm, bounce 11 cm; drop 40 cm, bounce 22 cm; drop 60 cm, bounce 32 cm; drop 80 cm, bounce 44 cm; drop 100 cm, bounce 53 cm.

Claim. For this ball, the bounce height is about a fixed fraction of the drop height — roughly half.

Evidence. The bounce-to-drop ratio is 11/20 = 0.55, 22/40 = 0.55, 32/60 ≈ 0.53, 44/80 = 0.55 and 53/100 = 0.53 — between 0.53 and 0.55 at every height tested.

Reasoning. If the bounce height grew in a different way from the drop height, the fraction would change from row to row. Instead it holds steady — about 0.54 on average — across all five heights, so the pattern the data supports is a roughly fixed fraction, not a rising or falling one.

Limits. The argument holds for this ball, this surface, and drops up to 100 cm only. The two lowest fractions (0.53) sit at the highest drops, so beyond 100 cm the fraction might start to fall — the data cannot say, because nothing was measured there.

Practice questions

Scaffolded questions

Question 1 Use the argument-parts figure (claim, evidence, reasoning, limits). (a) Name the three working parts of an evidence-based argument. (b) What job does the limits part do? (c) The evidence row in the figure quotes "about 1.0 g extra per 1 °C". Why is that better reasoning than writing "the results show the claim is true"?

Question 2 Use the primary/secondary figure. Classify each as primary or secondary data for the student using it, and give the reason. (a) The bounce-height trials your own class ran last week. (b) Bureau of Meteorology rainfall records downloaded for a project. (c) A news article describing a study published in a journal.

Question 3 Use the sugar-dissolving figure (21 g at 10 °C to 81 g at 70 °C). (a) Describe the trend in the data. (b) Calculate the average rise in dissolved mass per 1 °C across the full range. (c) Which conclusion can the data carry: *"the hotter the water, the more sugar dissolves"* or *"any solid dissolves better in hotter water"*? Give the reason.

Question 4 Use the headline-check figure (two simulated headlines, five checks). (a) For headline 1 ("Gene found for a common disease — a cure within five years"), which check catches the biggest problem, and what is the problem? (b) For headline 2 ("Carbon dioxide in the air hits a new record"), state one question check 2 would ask and one question check 3 would ask. (c) Check 5 asks who paid for the work. Why does funding matter even if the study itself was done carefully?

Question 5 Use the might-vs-cause figure (one survey of 300 students, three reports). (a) Which reports does the survey support — A, B, C, or all three? (b) Why can this survey not support report C? (c) What kind of investigation *could* test whether fizzy drinks cause later bedtimes? Name the condition that makes it fair.

Question 6 Use the citation figure (five-part source line). (a) Match each part of the source line — Bureau of Meteorology; "Climate statistics"; 2026; bom.gov.au; retrieved 21 August 2026 — to its job (who made it; title; year; where to find it; when you looked). (b) The red box quotes "Melbourne's rainfall is falling" with no source line. State two things a reader cannot do with that claim. (c) Why does a source line include the date you looked, not just the page?

Unscaffolded questions

Question 7 Use the evidence-chain figure of the continental-drift argument. The coastline fit and the Mesosaurus fossils both point towards joined continents, but they do different jobs in the argument. Explain why the Mesosaurus evidence is stronger than the coastline fit alone, and why four lines together are stronger than any one of them.

Question 8 An advertisement says: *"Most dentists recommend Sparkle toothpaste."* When asked, the company says it surveyed dentists at one conference, without stating how many or what question was asked. Using the reading-a-statistic questions (how many; who; how chosen; how asked), explain why the claim is weak as it stands, and write a version of the claim the company could fairly make from its own information.

Question 9 Two reports argue about whether most students want a salad bar at the canteen. Report 1: at a club meeting, 18 of the 25 students present voted yes. Report 2: in a school poll, 483 of 900 students voted yes. Calculate both percentages, state which report is the better basis for the claim "most students want a salad bar", and give two reasons.

Question 10 A student copies a paragraph from a website into her report, changes three words, and puts no source line. Name what is wrong (there is more than one thing), and rewrite the situation the way it should look: what she does with the words, and what she adds.

Question 11 A group wants to use images of Aboriginal rock art that they found on a public website, in a presentation about Australian landscapes. Use the ethics-flow figure and the cultural-protocols theory. (a) Which question in the ethics flow does this material raise? (b) State two things the group must check before using the images. (c) State what the group should do if they cannot confirm those things.

Question 12 A class investigates stirring and dissolving (simulated data for practice): the same mass of sugar in the same water, unstirred it takes 90 s to dissolve; slow stirring 55 s; medium stirring 38 s; fast stirring 27 s. Construct a complete argument — claim, evidence, reasoning, limits — including what the *savings* (the seconds each step of stirring saves) show about the pattern.

Question 13 An article says: *"Chocolate makes you cleverer. Researchers surveyed 50 students about how much chocolate they eat and their last test scores, and students who reported eating more chocolate had higher scores."* Run the five headline checks on this article. State what each check finds, and write the strongest claim the survey actually supports.

Question 14 Two sources disagree about a question for a project. Source A is the Museums Victoria website; source B is a personal blog written by someone whose name and qualifications are not stated. Using the six source questions, explain which source you would cite and why, naming at least three questions where the two sources differ.

Question 15 A company that sells vitamin drinks pays for a study, which finds that students who drink vitamin drinks report feeling more focused. The company's advertisement then says: *"Science proves vitamin drinks improve focus."* (a) Which headline check is most relevant to the funding, and what should the advertisement have declared? (b) The study was a survey asking students how they felt — explain why "improve" is too strong a word for what the study showed. (c) Write a fair one-sentence report of the study.

Question 16 A student wants to support the claim *"our creek is cleaner than last year."* She has her own data from one visit this year (water clarity and one invertebrate count) and a newspaper article saying the creek was polluted last summer. (a) Classify each source as primary or secondary for her. (b) State two of the six source questions she should run on the newspaper article. (c) Write the conclusion her evidence can actually carry, limits included.

Answers

Question 1 (a) claim, evidence, reasoning; (b) it states what the evidence cannot show — the conditions, substances or ranges the argument does not reach; (c) it points at the pattern in the numbers itself, so the reader can follow *why* the evidence supports the claim instead of being asked to take the writer's word. **Question 2** (a) primary — her class collected it and she is using it; (b) secondary — the Bureau collected it, she is accessing it; (c) secondary — the study's authors collected it; the journalist and the student are both downstream of the collection. **Question 3** (a) the dissolved mass rises at every temperature tested; (b) (81 − 21) g ÷ (70 − 10) °C = 60 ÷ 60 = 1.0 g per 1 °C; (c) the sugar one — only sugar was tested, so the data cannot speak for "any solid". **Question 4** (a) check 3 — "a cure within five years" is a prediction the study did not measure; the finding was a gene; (b) check 2: how many measurements, and over how long, is the record based on? check 3: is the record about the whole atmosphere, or about one measuring station?; (c) because a funder with a stake in the result may shape what is studied or reported — the funding does not make the study wrong, but readers are entitled to know. **Question 5** (a) A and B; (b) a survey only shows the two things vary together — students who drink more fizzy drinks may differ in other ways too, so the cause cannot be pinned on the drinks; (c) a fair test — the same students, same everything except the fizzy drinks (e.g. fizzy drink on one night, none on another, bedtimes recorded) — only the tested factor differs. **Question 6** (a) Bureau of Meteorology = who made it; "Climate statistics" = title; 2026 = year; bom.gov.au = where to find it; retrieved 21 August 2026 = when you looked; (b) cannot check who made it, and cannot find it to check it against the evidence; (c) a web page can change after you read it — the date tells the reader which version you saw. **Question 7** A coastline shape is suggestive but coastlines change, so the fit alone is only a hint; Mesosaurus was a freshwater reptile that could not cross an ocean, so its fossils on both sides of the Atlantic are hard to explain without joined land. Together the lines are stronger because they are independent — shape, fossils, rocks and ancient ice all point the same way, and one claim (joined continents) explains all of them at once. **Question 8** The sample is unknown in size (how many?), chosen from one conference only (who? — conference attendees may already favour the brand or its sponsors), and the question asked is unstated (how asked? — "recommend" could mean "recommend at all" or "recommend most"). A fair claim: "At one dental conference, the dentists we surveyed recommended Sparkle" — nothing broader until the sample and question are stated. **Question 9** Report 1: 18 ÷ 25 = 72%; Report 2: 483 ÷ 900 = about 54%. Report 2 is the better basis: far more students (900 vs 25), and the whole school rather than one club's meeting — students who attend that club may already favour salads. Even Report 2 only says a bare majority: "most" is fair, "overwhelmingly" is not. **Question 10** Two things wrong: the words are still someone else's (changing three words does not make them hers), and there is no acknowledgement. Fix: use her own words to say what the source showed, keep only what she needs, and add a source line — who made it, title, year, where, when she looked. Copying the exact words would need quote marks *and* the source line. **Question 11** (a) "Does it touch Aboriginal or Torres Strait Islander cultural heritage or knowledge?" — yes; (b) any two of: whether the website has permission to share the images; the cultural protocols and permissions that apply to that material (e.g. whether it may be shared publicly); the terms of use of the website itself; (c) don't use the images — ask the teacher, and check the relevant published instrument or source, before anything is used. **Question 12** Claim: stirring shortens the time to dissolve, but each step up in stirring speed saves less than the one before. Evidence: 90 s → 55 s → 38 s → 27 s; savings of 35, 17 and 11 s. Reasoning: every stirring speed beats the one before it, so stirring helps; but the shrinking savings (35 then 17 then 11) show the help is running out — faster and faster stirring buys less and less time. Limits: this sugar, this water, these four speeds only; what "even faster" stirring would do is not measured. **Question 13** Check 1: where was it done and is it published — unknown from the article. Check 2: only 50 students — a modest sample. Check 3: the article says "cleverer", but the study measured last-test scores — the headline overstates what was measured. Check 4: a survey shows association only, so "makes" is a cause-word the study cannot support. Check 5: who paid — not stated. Strongest claim the survey supports: "In this survey of 50 students, students who reported eating more chocolate also had higher last-test scores." **Question 14** Source A (Museums Victoria): question 1 — a museum with staff and collections behind it, strong on natural-history questions; question 2 — its pages state evidence you can check; question 4 — its claims can be checked against other institutions. Source B: question 1 — no stated author or qualifications; question 2 — no evidence shown; question 5 — unknown. Cite source A; the blog fails too many checks to carry a claim. **Question 15** (a) Check 5 — the advertisement should declare that the seller of the drinks paid for the study; (b) the study asked students how they felt — self-reported focus, measured by a survey — so it can show at most an association between vitamin drinks and *reported* focus; "improve" claims a cause and a measured change, neither of which the study produced; (c) e.g. "In a study paid for by the drink company, students who drank vitamin drinks reported feeling more focused; the survey cannot show the drinks caused it." **Question 16** (a) her own data — primary; the newspaper article — secondary; (b) any two of: who wrote it and why are they a good source on water quality? does it show its evidence (measurements), so it can be checked? how old is it — "last summer" against her one visit this year? who paid?; (c)

e.g. “On my one visit this year, the creek’s clarity and invertebrate count were consistent with a cleaner creek than the newspaper reported for last summer; with one visit and one secondary report, the claim needs repeat visits and direct measurements from both years before it can stand.”

Fully-worked solutions

Question 1 (a) The claim (what you argue), the evidence (what was observed or measured), and the reasoning (why the evidence supports the claim). (b) The limits state what the evidence *cannot* show — the substances, conditions or ranges the argument never reached — so the claim is not read more widely than the evidence allows. (c) Because reasoning must connect the evidence to the claim: quoting the pattern (about 1.0 g extra at each 1 °C) shows the reader the link, while “the results show it” asks the reader to trust the writer without seeing the thinking.

Question 2 (a) Primary data: her class collected it, and she is one of the people using it — collected by the person using it is the definition. (b) Secondary data: the Bureau of Meteorology collected the rainfall records; the student is accessing them for her own work. (c) Secondary data — twice over. The journal’s authors collected the study’s data, and the news article is already a second-hand account of it; the student is one step further along the same route.

Question 3 (a) The mass dissolved rises at every temperature tested: 21 → 36 → 52 → 66 → 81 g. (b) Rise in mass = 81 − 21 = 60 g; rise in temperature = 70 − 10 = 60 °C; average rise = 60 ÷ 60 = 1.0 g per 1 °C. (c) Only *“the hotter the water, the more sugar dissolves.”* The investigation tested one substance — sugar — so the data can only speak for sugar; “any solid” reaches past everything that was measured.

Question 4 (a) Check 3, “does the headline match what was actually measured?” The study’s finding was a gene; the cure, and its five-year timeline, is a prediction that was not measured. (b) Check 2: how many measurements, at how many places, over how long, is the record based on? Check 3: does “the air” mean the whole atmosphere, or the record at one measuring station? (c) A funder with a stake in the answer may shape what gets studied, how, or what gets reported. None of that makes the study wrong — but check 5 exists because readers are entitled to know who paid before they weigh the result.

Question 5 (a) Reports A and B. Both stay inside what the survey found: the two things vary together. (b) Because the survey tested nothing — it only asked students what they did. Students who drink more fizzy drinks may also differ in screen time, sport, age or anything else, and any of those could sit behind the later bedtimes; a survey cannot rule them out. (c) A fair test: the same students under the same conditions except for the fizzy drinks (with fizzy drink one night, without another, bedtimes recorded both times). The condition that makes it fair is the one from Subtopic 3A: only the tested factor differs.

Question 6 (a) Bureau of Meteorology — who made it; “Climate statistics” — the title; 2026 — the year; bom.gov.au — where to find it; retrieved 21 August 2026 — when you looked. (b) The reader cannot say who made the claim, and cannot find the evidence to check it — the two things a source line exists to make possible. (c) Because a web page can change after you read it. The retrieval date tells the reader which version of the page your claim was based on.

Question 7 The coastline fit alone is weak because coastlines change — erosion, sea level and sediment reshape them — so a matching shape is suggestive, not proof. The Mesosaurus evidence does more work: Mesosaurus was a freshwater reptile, so it could not have crossed a salt ocean; its fossils on both sides of the Atlantic need joined land, or an explanation just as large. The four lines together beat any one of them because they are *independent*: coastline shape, fossils, rocks and mountain chains, and ancient-ice scratch marks have nothing to do with each other except where they point. When independent lines converge, and one claim — joined continents — explains all of them at once, the claim earns trust no single line could give it.

Question 8 How many — the company does not say; ten dentists and five hundred dentists are different claims. Who — dentists at one conference, who may favour the brand or its sponsors, not dentists in general. How chosen — no method stated, so the sample may be hand-picked. How asked — the exact question is unstated; “do you recommend Sparkle?” answered once is not “most dentists’ first choice”. A fair claim stays inside what was actually done: *“At one dental conference, the dentists we surveyed recommended Sparkle toothpaste.”* Anything broader needs a stated sample, method and question behind it.

Question 9 Report 1: 18 ÷ 25 = 72%. Report 2: 483 ÷ 900 ≈ 54%. Report 2 is the better basis for "most students want a salad bar", for two reasons. First, the sample: 900 students across the school beats 25 students at one club meeting, and students who attend that club may already favour salads — Report 1's sample is drawn from a corner of the students it claims to describe. Second, the honesty of the size: 72% sounds like a landslide, but it is the small poll's number; the big poll says a bare majority. So: most students want a salad bar — fairly said — but neither poll says much more than that.

Question 10 Two separate failures. First, the words are still the website's: changing three words keeps the sentence's shape and content, which is the website's work, not hers. Second, there is no acknowledgement, so a reader has no way to see what is hers and what is not. The fix: she writes what the source showed *in her own words*, keeps only what her report needs, and adds a full source line — who made it, title, year, where to find it, when she looked. If she truly needs the exact words, they go in quote marks, with the source line either way.

Question 11 (a) The third question of the ethics flow: *"Does it touch Aboriginal or Torres Strait Islander cultural heritage or knowledge?"* Rock art is Aboriginal cultural heritage, so the answer is yes — the flow goes to the cultural-protocol branch, not straight to "OK to use". (b) Any two of: whether the website has permission to share those images in the first place; what cultural protocols and permissions apply to that material — whether it may be shared publicly, and with whom, is decided by its owners; and the website's own terms of use. (c) The group does not use the images. The rule for uncertainty in this subtopic is don't use it — ask: their teacher, and the relevant published instrument or source, before anything is presented.

Question 12 Claim: stirring shortens the dissolving time, but each step up in stirring speed saves less time than the one before. Evidence: unstirred 90 s; slow 55 s; medium 38 s; fast 27 s. The savings are 90 − 55 = 35 s, 55 − 38 = 17 s, and 38 − 27 = 11 s. Reasoning: every stirring speed beats the one before it (55 < 90, 38 < 55, 27 < 38), so stirring clearly helps. But the savings shrink — 35, then 17, then 11 — so the help is running out: each extra step of speed buys less time than the last. Limits: this sugar, this volume of water, and these four speeds only. Nothing was measured past "fast", so whether even faster stirring would save anything at all — or whether the time would stop falling — this data cannot say.

Question 13 Check 1 — where was it done, and is it published? The article does not say. Check 2 — 50 students: a modest sample for a claim about cleverness. Check 3 — the headline says "cleverer", but what was measured was last-test scores; the headline overstates the measurement. Check 4 — the study was a survey, so it shows an association at most; "makes" is a cause-word the study cannot support. Check 5 — who paid is not stated. The strongest claim the survey actually supports: *"In this survey of 50 students, students who reported eating more chocolate also had higher last-test scores"* — with the cause still open.

Question 14 Cite source A. Run the six questions: question 1 — Museums Victoria is a museum with staff, collections and a public role behind it, a good source on a natural-history question, while the blog states no author and no qualifications at all. Question 2 — the museum's pages state evidence a reader can check; the blog offers an assertion. Question 4 — the museum's claim can be checked against other institutions working separately; the blog has no independent standing to check against. (Question 5 cuts the same way: the blog's backer is unknown.) Where sources disagree, the one that survives the checklist carries the claim — and where the disagreement matters, the honest report says so: "Source A says this; an unsigned blog says otherwise."

Question 15 (a) Check 5 — who paid for the work. The advertisement should have declared that the study was paid for by the company that sells the drinks. (b) Because the study was a survey of self-reported focus: students said how they felt. That can show, at most, that vitamin-drink drinkers *reported* feeling more focused — an association. "Improve" claims a measured change caused by the drink, and the study measured neither a change nor a cause. (c) Example: *"In a study paid for by the company that sells the drink, students who drank vitamin drinks reported feeling more focused; because the study was a survey, it cannot show the drinks caused the reported focus."*

Question 16 (a) Her own clarity and invertebrate data — primary data (she collected it). The newspaper article — secondary data (someone else collected and reported it). (b) Any two of the six: who wrote the article, and why is that person a good source on water quality? Does the article show its evidence — measurements, not just the word "polluted" — so it can be checked? How old is the report, and does "last summer" line up with her visit this year? Is it declared who paid for any study behind it? (c) Example conclusion with limits: *"On my one visit this year, the creek's clarity and invertebrate count were consistent with a cleaner creek than the newspaper reported for last summer; but one visit this year and one secondary report from last year are not yet a comparison — repeat visits, and direct*

measurements from both years, are needed before the claim can stand." That is the shape of a conclusion the evidence can carry: it says what the data showed, and it says exactly what is still missing.

Where this leads: the reporting steps of the inquiry sequence (Subtopics 6A and 6B) take over from here — the argument, the source lines and the permissions you checked in this subtopic are exactly what a report must show. Subtopic 5A judged the method; this one built the case from it, and assessed everyone else's. The ethics thread also continues later in the book: how funding and priorities shape what science gets done — the question behind check 5 — is where Chapter 16's look at science in society begins.