Writing Classes for Mathematical Modeling Competitions

W3: Building Trust – Assumptions, Validation, and Sensitivity

Kenneth Sok Kin Cheng

Revised 20 August 2026

Learning goals

By the end of this two-hour writing class, students should be able to:

- explain how assumptions, verification, validation, sensitivity, and uncertainty play different roles in a modeling report;
- write a four-part **assumption card**: statement, rationale/evidence, model consequence, and stress test;
- distinguish a modeling assumption from a definition, parameter value, scope choice, or mathematical requirement;
- build an **assumption ledger** that links each assumption to an equation, output, and test;
- distinguish **verification** from **validation**, and distinguish both from calibration;
- select validation evidence appropriate to the model's purpose and available data;
- design one-at-a-time, scenario, structural, and simulation-based sensitivity checks;
- report sensitivity using input ranges, output metrics, decision changes, and operational triggers;
- separate parameter sensitivity from uncertainty in estimated outputs;
- write a concise trust section that supports rather than merely praises a recommendation.

1 Trust is an evidence structure, not a confidence adjective

A model becomes useful only when a reader can see why its conclusions deserve attention. A sentence such as "our model is accurate and robust" adds no trust by itself. Trust comes from a visible chain linking the real-world interpretation to tests and decision consequences.

Official competition guidance supports this emphasis. Current COMAP guidance asks teams to state assumptions with justifications and to analyse their work, including sensitivity to changed or removed assumptions. IM^{2}C resources describe modeling as a cycle in which solutions are tested and evaluated against the real situation. Exact section names may vary, but the evidence obligations do not.

Launch activity: rank the trust claims

A team recommends four buses. Rank the following claims from least to most useful, then explain what evidence is missing.

1. "Our model is highly reliable."
2. "All buses satisfy capacity 60 in the baseline data."
3. "Predicted route times have mean absolute error 2.4 minutes on 12 held-out school days."

4. “Four buses remain sufficient in 18 of 20 demand scenarios; a fifth is required when total ridership exceeds 240.”

The last three answer different questions. None replaces the others.

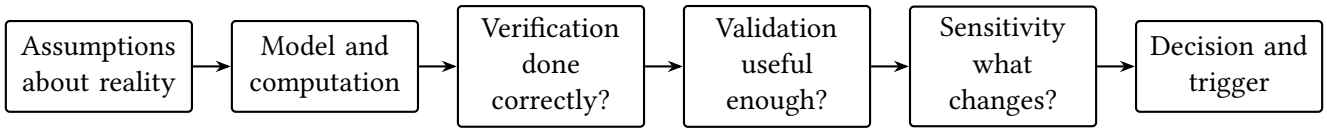

Figure 1: A report earns trust through a chain of evidence. A weakness in one link should appear as a qualification or further test, not be hidden.

Five terms that must not be merged

Term	Question	Typical evidence
Assumption	What claim about the system makes the model possible?	Domain evidence, data pattern, scope rationale, and a stated trade-off.
Calibration	What parameter values make the model fit the chosen data?	Estimation procedure, training data, objective function, fitted values.
Verification	Did we implement and solve the stated model correctly?	Units, invariants, hand cases, code tests, convergence, independent implementation.
Validation	Is the model adequate for its intended purpose?	Held-out data, external comparison, field evidence, cross-method prediction, domain plausibility.
Sensitivity / uncertainty	How do inputs and imperfect knowledge affect outputs or decisions?	Perturbations, scenarios, distributions, intervals, structural alternatives, decision switches.

Checkpoint: Checkpoint: classify before you write

Classify each sentence.

1. “We estimate service rate μ from 40 observed service times.”
2. “When arrivals are set to zero, the simulated queue remains empty.”
3. “Predicted average wait differs from observed wait by 0.8 minutes.”
4. “Replacing exponential service with fixed service reduces the recommendation from three servers to two.”
5. “We treat morning demand as stationary from 7:15 to 8:00.”

Suggested labels: calibration, verification, validation, structural sensitivity, assumption.

2 Write assumptions as testable modeling choices

An assumption is neither an apology nor a list of obvious facts. It is a deliberate claim about the system that changes the mathematics or the scope of the conclusion.

The four-part assumption card

For every material assumption, write:

1. **Statement – what?** A precise claim that could be false.
2. **Rationale/evidence – why?** Data, literature, domain logic, or a transparent simplification reason.
3. **Model consequence – effect?** Which equation, algorithm, parameter, or computational burden changes because of it?
4. **Risk/stress test – then what?** How might the assumption fail, and which sensitivity or validation check addresses that risk?

The fourth part closes the loop between the Assumptions section and later evidence.

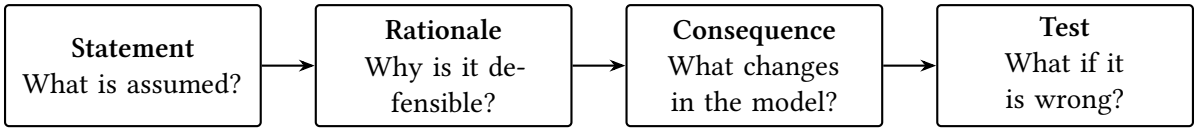

Figure 2: A complete assumption card connects interpretation, mathematics, and later testing.

Worked example 1: assumption cards for the school-bus case

Assumption A – registered riders represent daily demand.
Statement. The baseline routing problem includes all 226 registered riders and treats each as travelling on the modeled morning.
Rationale. Registration data are the only complete location data available and provide a conservative ordinary-day demand estimate.
Model consequence. The capacity constraints use 226 fixed demand points, producing the lower bound $\lceil 226/60 \rceil = 4$ buses.
Risk/test. Absence and late-registration patterns can change district loads. We therefore rerun the model under 20 district-demand scenarios with changes of up to ± 10 riders.

Assumption B – observed morning travel times are reusable.
Statement. Link travel times estimated from 12 normal school mornings are used for future ordinary weekdays.
Rationale. The data exclude holidays, roadworks, and severe-weather days, and the within-link variation is smaller than the between-link variation.
Model consequence. Routes can be evaluated through a fixed time matrix rather than a time-dependent traffic simulator.
Risk/test. Peak congestion may increase all route times. We test multiplicative travel-time factors from 0.90 to 1.10 and state a delay trigger.

Model criticism: Do not force every sentence into the Assumptions section

- "Bus capacity is 60" may be an input or policy constraint, not an assumption, if it is specified by the problem.
- "Let $x_{ij} = 1$ when bus i serves stop j" is a definition.
- "Every route must begin and end at school" is a feasibility rule.
- "Travel times are stable across ordinary mornings" is an assumption because reality may violate it.
- "We only study the morning journey" is a scope decision; explain it where the scope is introduced and record its effect on conclusions.

The section should contain choices that matter, not every fact used in the report.

The assumption ledger

A ledger prevents assumptions from disappearing after page 3.

Assumption	Where it enters	Risk if false	Evidence or test
Daily demand $= 226$	Capacity and route allocation	Overload or unnecessary fleet	Demand scenarios; threshold at 240 riders
Stable link times	Time matrix and ride limit	Underestimated maximum ride	Holdout GPS error; 0.90–1.10 time factors
No road closure	Shortest-path network	Infeasible planned route	Scenario removing the most-used road
One pickup cluster per stop	Demand aggregation	Extra walking burden	Maximum walking-distance check

An assumption with no downstream effect is probably irrelevant. A high-risk assumption with no test is a visible evidence gap.

Checkpoint: Assumption audit

For one assumption in your current project, underline the exact equation, data-processing step, or algorithm that depends on it. Then write the sentence that will report its stress test. If neither can be identified, revise or remove the assumption.

3 Verification: did we solve the stated model correctly?

Verification concerns the mathematics and implementation of the model as written. A model can be perfectly verified and still be invalid for reality; it can also be conceptually suitable but incorrectly coded.

A practical verification menu

1. **Dimensional/unit check:** both sides of every equation have compatible units.
2. **Constraint check:** every reported solution satisfies capacity, non-negativity, budget, conservation, and logical rules.
3. **Limiting or hand case:** a tiny case with a known answer is reproduced.
4. **Invariant check:** conserved totals remain conserved; probabilities sum to one; flows balance.
5. **Numerical refinement:** reducing step size or tightening tolerance changes outputs negligibly.
6. **Independent implementation:** a second method, package, or teammate reproduces key results.
7. **Reproducibility record:** data version, random seed, solver settings, and stopping rules are stated when material.

Choose checks that match the model. Do not write a generic paragraph claiming that "the code was checked."

Worked example 2: verification paragraph

Before interpreting the optimized routes, we verified feasibility and implementation. All 226 students are assigned exactly once, no bus exceeds capacity 60, every route starts and ends at school, and computed route distance equals the sum of its constituent links. On a six-stop hand case, the program reproduces the enumerated optimum of 18.6 km. Running the local-search phase from 50 random initial solutions produces the same best objective in 43 runs and values within 0.7% in all runs. These checks support the claim that the reported solution is a consistent output of the stated routing model; they do not yet establish that the travel-time assumptions represent future mornings.

Model criticism: Calibration is not validation

A fitted regression can match its training data because its parameters were chosen to do so. A queue model can reproduce a sample mean after its service rate is calibrated to that mean. This establishes calibration, not independent validation.
Reserve some evidence for testing: a later time period, held-out locations, external cases, a different observable, or an independent method. When data are scarce, state the limitation rather than calling in-sample fit "validation."

4 Validation: is the model adequate for the decision?

Validation is purpose-dependent. A route-time model may be adequate for choosing between two route plans even if it cannot predict every individual journey to the nearest minute. State the purpose, metric, comparator, and acceptance logic.

The validation claim template

A complete validation claim answers four questions:

1. **Target:** What observable or decision-relevant quantity is tested?
2. **Comparator:** Against what data, benchmark, alternative method, or established range?
3. **Metric:** Error, agreement rate, rank correlation, coverage, or qualitative pattern?
4. **Judgment:** Why is this level of agreement adequate for the model's intended use?

Example: "On 12 held-out mornings, route-time predictions have MAE 2.4 minutes and maximum error 5.1 minutes. Because route comparisons differ by 8–19 minutes, this accuracy is adequate for ranking the candidate plans, although not for promising exact arrival times."

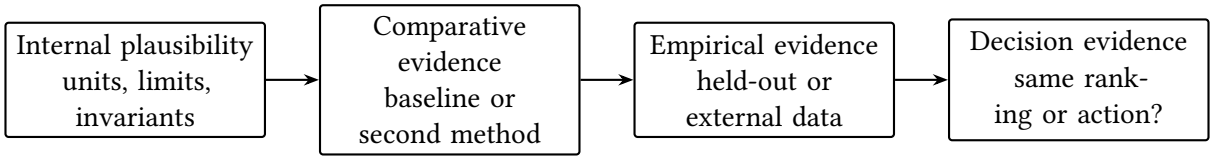

Figure 3: Validation strength usually increases as evidence becomes more independent and more closely tied to the intended decision. Not every project can reach every level.

A model-specific validation menu

Model type	Useful validation evidence	Weak substitute to avoid
Regression / prediction	Held-out error, residual pattern, baseline comparison, interval coverage	Reporting only training R^2
Optimization	Feasibility, lower/upper bounds, comparison with baseline or benchmark, scenario performance	Calling the solver output "optimal" without gap or scope
Simulation	Hand case, analytical special case, convergence across replications, output-data comparison	One run with no random-seed or uncertainty report
Dynamic model	Initial-condition reproduction, conservation, qualitative trend, out-of-sample trajectory	A visually similar curve fitted to the same data
Decision index / ranking	Expert or external ranking, stability of top choices, back-test on known cases	High internal consistency with no external criterion

Worked example 3: validation with a benchmark and held-out data

The routing model is validated at two levels.
Travel-time prediction. Twelve school mornings are held out from model construction. Across the 48 route-day combinations, the mean absolute error is 2.4 minutes and the largest error is 5.1 minutes. Errors show no systematic increase with route length.
Decision comparison. The model ranks the existing route plan worse than the proposed plan on all 12 mornings. The observed mean difference is 13.6 minutes per rider, compared with a predicted difference of 14.7 minutes. The recommendation therefore survives the validation data even though exact route times remain uncertain.

Checkpoint: Repair the validation sentence

Weak: "Our model agrees closely with the real data and is therefore valid."
Rewrite it by adding the target, comparator, metric, and decision judgment. Avoid the absolute phrase "the model is valid"; validation supports adequacy for a stated purpose, not universal truth.

5 Sensitivity and uncertainty: will the decision survive change?

Sensitivity analysis varies inputs or modeling choices and measures their effect. **Uncertainty analysis** represents imperfect knowledge about inputs or stochastic outputs and reports a range or distribution. They are related but not identical.

Four useful levels of sensitivity analysis

1. **Local/OAT sensitivity:** vary one input over a plausible range while holding the others fixed. Fast and interpretable, but misses interactions.
2. **Scenario analysis:** vary several inputs together in coherent situations such as ordinary, peak, and disruption days.

3. **Structural sensitivity:** replace a distribution, objective, functional form, aggregation rule, or model family.
4. **Global/simulation-based sensitivity:** sample uncertain inputs jointly and study output distributions or importance measures.

The method should match the decision. A small two-parameter model may need only OAT plus a structural check; a nonlinear stochastic system may need joint sampling.

Report sensitivity in decision language

For each varied input, state:

1. nominal value and plausible range;
2. output metric and baseline value;
3. magnitude or direction of output change;
4. whether the recommendation, ranking, feasibility, or threshold changes;
5. what the finding implies for data collection or contingency planning.

A result can be numerically sensitive but decision-stable, or numerically stable but close to a hard constraint. Report both the output and the action.

Worked example 4: OAT table and decision boundary

Input/test	Baseline	Lower case	Upper case	Decision effect
Total ridership	226	206	246	Four buses below 241; five at 246
Travel-time factor	1.00	0.90	1.10	Distance plan unchanged; 45-min limit fails above 1.02
Bus capacity	60	54	66	Five buses at 54; four at 60 or 66
Distance weight in objective	1.00	0.75	1.25	Same fleet; route order changes slightly
Road availability	all links	–	remove busiest link	Four buses remain feasible; distance rises 6.8 km

The fleet recommendation is stable to moderate travel-time and network changes but changes at two clear thresholds: ridership above 240 or effective capacity below 57. The ride-time service standard is more fragile than the fleet size and therefore needs a scheduling buffer.

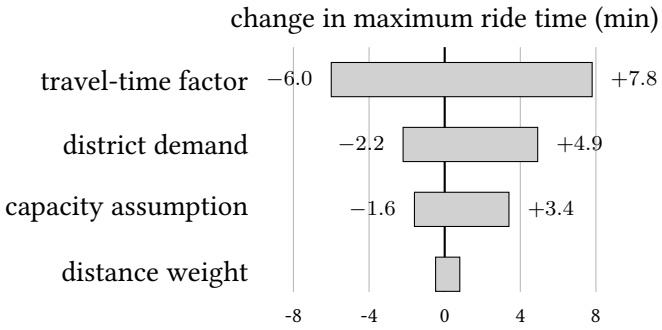

Figure 4: Illustrative tornado diagram. Bars show output change over stated input ranges; they do not by themselves show probability or decision thresholds.

Model criticism: A tornado diagram needs a table behind it

A bar chart without stated input ranges is not reproducible. A sorted bar chart also does not show parameter interactions or likelihood. Include the nominal values and ranges in a table or caption, then use prose to identify the decision-relevant threshold.

Optional normalized sensitivity

When inputs have different units, a local **elasticity** can help compare relative influence:

$$E_{y,x} \approx \frac{\Delta y / y_0}{\Delta x / x_0}.$$

An elasticity of 1.5 means a 1% input change produces approximately a 1.5% output change near the baseline. This is a local summary, not proof of global linearity.

Worked example 5: structural sensitivity changes the recommendation

The shortest-total-distance formulation produces an 82.0-km plan, but one student group remains on a bus for 51.0 minutes. Replacing the objective with a balanced objective that penalizes maximum ride time gives 87.4 km and a 44.1-minute maximum.
The recommendation is therefore structurally sensitive to the objective function. We do not call the 82.0-km plan "better" merely because its total distance is lower. We recommend the balanced plan because it satisfies the stated service standard, while reporting the 5.4-km efficiency cost of that choice.

6 Integrated case: write the complete trust story

The following example reuses the school-bus evidence from W2. The purpose is to show how the same numerical work becomes three distinct report sections.

Worked example 6: competition-ready Assumptions paragraph

A1: ordinary-day demand. We model all 226 registered riders as travelling on the target morning. Registration data provide the only complete geocoded demand set and give a conservative baseline. This assumption fixes the capacity requirement and implies a minimum of four 60-seat buses. Because absence and new registration can redistribute loads, Section 7 tests 20 district-demand scenarios and reports the fleet-change threshold.
A2: reusable travel times. Link times are estimated from 12 normal school mornings and treated as representative of future ordinary weekdays. This avoids a separate traffic-flow simulation and allows

candidate routes to be compared through one time matrix. The assumption is weaker on disruption days, so validation uses held-out mornings and sensitivity multiplies all link times by factors from 0.90 to 1.10.

Worked example 7: competition-ready Verification and Validation paragraph

Verification. Every rider is assigned exactly once; capacities, school start/end requirements, and route continuity are satisfied. The program reproduces a six-stop enumerated optimum and returns the same best objective in 43 of 50 random starts.
Validation. On 12 held-out mornings, predicted route times have MAE 2.4 minutes and maximum error 5.1 minutes. The model predicts that the proposed plan improves mean ride time by 14.7 minutes; the observed improvement is 13.6 minutes. This agreement is sufficient for comparing plans, although individual arrival times should retain a buffer.

Worked example 8: competition-ready Sensitivity and Decision paragraph

Demand perturbations of up to ± 10 riders per district leave the four-bus plan feasible in 18 of 20 scenarios. The two failures occur when total ridership exceeds the aggregate capacity of 240. Travel-time perturbation does not change fleet size, but the 45-minute service limit is crossed when ordinary-day link times rise by more than approximately 2%. Removing the most-used road increases distance by 6.8 km but does not change the fleet recommendation.
We therefore recommend four buses for ordinary days, a three-minute timetable buffer, and an early review at 236 projected riders and a fifth-bus trigger from 241 confirmed riders or a major disruption is forecast. This is a conditional operational recommendation, not a claim that four buses are universally sufficient.

Why the three sections should not repeat each other

- **Assumptions** explain the choices and identify risks before results are presented.
- **Verification/validation** show whether the implemented model and its predictions are credible.
- **Sensitivity/uncertainty** show where the output or decision changes and convert that boundary into a trigger.

Repeating "the model is reasonable" in all three sections wastes space and leaves the evidence invisible.

7 Writing patterns that weaken trust

Model criticism: Repair these common overclaims

Weak wording	Evidence-based repair
"The assumption is reasonable."	State the source, observed range, simplification benefit, and failure mode.
"The model is validated."	State the tested target, data or benchmark, metric, and intended-use judgment.
"The result is robust."	State the input range, output change, and whether the decision changed.
"The error is small."	Give the metric, units, baseline scale, and comparator.
"Monte Carlo confirms the model."	State what was sampled, how many replications, uncertainty, and which analytical claim was reproduced.
"Changing assumptions has little effect."	Name the changed assumption and quantify the change in output or decision.

Checkpoint: The evidence sentence test

Circle every adjective in your current trust sections: reasonable, accurate, realistic, reliable, robust, stable, negligible, significant. For each, add the evidence needed to make it measurable, or delete it.

8 Differentiated writing studio

Core route: build the assumption ledger

For a cafeteria queue model, complete four ledger rows using these candidate assumptions: stationary arrival rate, independent arrivals, fixed number of servers, and no customer abandonment. For each row, identify where it enters, the risk if false, and one stress test.

Challenge route: design the validation plan

A regression predicts bicycle usage from temperature, rainfall, day type, and route length. You have 180 daily observations. Propose:

1. one calibration procedure;
2. two verification checks;
3. two validation metrics on held-out data;
4. one baseline model;
5. one criterion for whether the model is adequate for staffing a bicycle-storage facility.

Do not use training R^2 as your only evidence.

Extension route: turn uncertainty into a decision rule

A reservoir model predicts that a release of 24 ML/day keeps storage above the safety threshold in 91 of 100 sampled climate-demand scenarios. At 22 ML/day, success rises to 98 of 100 but water supply to farms falls by 8%. Write a recommendation that states the trade-off, uncertainty, and trigger for switching policies.

Peer audit: no author explanation

Exchange drafts and answer:

1. Which assumption creates the greatest decision risk?
2. What was verified, and what was validated?
3. Is any calibration evidence incorrectly called validation?
4. Which sensitivity range is most decision-relevant?
5. At what threshold does the recommendation change?
6. Which sentence overclaims beyond the evidence?

A trust section must survive without the author explaining what they intended.

9 Exit ticket and preparation for W4

Exit ticket

Write four sentences for your current project:

1. one complete assumption card compressed into one or two sentences;
2. one verification claim with a concrete check;
3. one validation claim with a comparator and metric;
4. one sensitivity claim ending with the condition that changes the decision.

If the four sentences sound interchangeable, the distinctions are not yet clear.

Class W3 take-aways

1. Assumptions should be connected to equations, risks, and later tests.
2. Verification asks whether the stated model was solved correctly; validation asks whether it is adequate for a stated purpose.
3. Calibration uses data to set parameters and should not be presented as independent validation.
4. Sensitivity analysis needs an input range, output metric, and decision consequence.
5. Structural sensitivity often reveals more than repeatedly perturbing numerical parameters.
6. Uncertainty should become an interval, probability, scenario frequency, or operational trigger.
7. Trust is earned by traceable evidence, not by adjectives.

In **W4: Academic English for Modeling Reports**, we move from evidence architecture to sentence and paragraph control: claim strength, hedging, active voice, equation integration, transitions, and common translation patterns.

Appendix A: Assumption-ledger worksheet

Assumption statement	Rationale/evidence	Model consequence	Risk and planned test

Appendix B: Validation-plan matrix

Claim to test	Independent evidence	Metric	Comparator	Adequacy rule

Appendix C: Sensitivity-report sentence frames

- *Varying [input] from [lower] to [upper] changes [output] from [value] to [value], while [decision] remains unchanged.*
- *The recommendation switches from [A] to [B] when [input] crosses approximately [threshold].*
- *Among the tested inputs, [parameter] has the greatest effect on [metric]; a $\pm X\%$ change produces [output range].*
- *Replacing [model assumption] with [alternative] changes [output/decision], indicating structural sensitivity to [feature].*
- *Across [number] jointly sampled scenarios, [decision] is feasible in [number or percentage]; failures occur primarily when [condition].*
- *The numerical output is sensitive, but the ranking remains stable because [margin or threshold].*

Appendix D: Before calling a result robust

Check that the report states:

☐ the input or assumption varied;

☐ the plausible range and its source or rationale;

☐ the output metric and baseline;

☐ the numerical change in the output;

☐ whether feasibility, ranking, or recommendation changed;

☐ interactions or structural alternatives considered, where material;

☐ the operational threshold or contingency implied by the test.

Appendix E: Official resources used in this lesson

- COMAP HiMCM/MidMCM Tips Article: emphasizes assumptions with justifications, clear modeling processes, sensitivity, strengths/weaknesses, conclusions, and readable presentation.
- COMAP problem guidance example: describes sensitivity as adjusting or removing an assumption to examine its impact.
- IM^{2}C Mathematical Modelling Framework: presents modeling as a cyclic process involving interpretation, mathematical work, evaluation, and refinement.
- IM^{2}C Writing a Modelling Report: stresses clear, complete communication so others can evaluate and use the model.

Contest-specific requirements can change. Re-check the current official rules and problem instructions before each competition.

Model audit card

Audit before trusting or communicating a model

Purpose	What question or decision does the model support?	**Data**	Which inputs are observed, derived, or simulated?
Assumptions	Which simplifications may change the answer?	**Units**	Are dimensions and scales consistent?
Verification	Were equations, code, constraints, and limiting cases checked?	**Validation**	Is performance tested against independent evidence?
Uncertainty	How do measurement, parameter, solver, and structural errors affect conclusions?	**Ethics**	Who benefits, who bears risk, and what fairness criterion matters?
Reproducibility	Are data provenance, versions, seeds, and transformations recorded?	**Communication**	Are claims conditional, traceable, and decision-relevant?

Further reading

- GAIMME: Guidelines for Assessment and Instruction in Mathematical Modeling Education (SIAM / COMAP)
- HiMCM / MidMCM Contest Instructions (COMAP)

Sources, provenance, and licence

These notes follow the iterative modeling and assessment perspective in the SIAM/COMAP GAIMME report. Competition-specific remarks are a snapshot checked on 20 August 2026 against the official COMAP HiMCM instructions; readers should verify current rules before submission. Unless a source is explicitly identified, classroom datasets and numerical scenarios are synthetic or simulated teaching examples and are not empirical claims about a real institution. Third-party references retain their own terms.

Student response guide and checking criteria

For closed calculations, show the definition, substitution, units, and at least one independent check; the worked examples in this unit are method hints, but you must complete the arithmetic yourself. Open modeling responses are checked on five dimensions: a clear purpose and output; stated data and provenance; defensible assumptions and scope; traceable computational verification and model validation; and alignment among uncertainty, limitations, and the recommendation. Support each dimension with concrete evidence rather than writing only "completed."