

Supplementary Appendix S1. Survey Instrument and Scoring Procedures

Lesson A and Lesson B post-lesson survey questions used identical wording. No participant identifying information or participant-generated identification codes are reproduced in this supplement.

Procedural Items

A participant-generated code field was used to generate codes for each participant's response.

The code generation recipe has been omitted here to protect against potential identification of participants in the relatively small classroom sample.

Timing check: "It has been AT LEAST 24 hours since my last video lesson." Response options included Yes, No, and This is my first video.

Supplementary Table S1. Exact Survey Items and Scoring Rules

Construct	Exact item wording	Response scale	Scoring
Effort	I put a lot of effort into this.	1–7: Not at all true to Very true	Mean of three items.
Effort	I tried very hard on this lesson.	1–7: Not at all true to Very true	Mean of three items.
Effort	I did not put much energy into this.	1–7: Not at all true to Very true	Reverse-scored as 8 minus response, then averaged.
Negative affect	Learning this lesson made me irritated.	1–5: Strongly disagree to Strongly agree	Mean of irritation, boredom, and shame.
Positive affect	I feel confident that I will be able to master the material of this lesson.	1–5: Strongly disagree to Strongly agree	Mean of confidence and enjoyment.
Negative affect	This material is so boring I find myself daydreaming.	1–5: Strongly disagree to Strongly agree	Mean of irritation, boredom, and shame.
Negative affect	I feel ashamed when I think I lack ability within this lesson.	1–5: Strongly disagree to Strongly agree	Mean of irritation, boredom, and shame.
Positive affect	I enjoy dealing with the lesson material.	1–5: Strongly disagree to Strongly agree	Mean of confidence and enjoyment.
Task clarity	My instructor made sure I really understood the goals of the course and what I needed to do.	1–7: Strongly disagree to Strongly agree	Mean of two task-clarity items.
Task clarity	I understood what I was supposed to do at each stage of the investigation.	1–7: Strongly disagree to Strongly agree	Mean of two task-clarity items.
Perceived emotional support	I feel that my instructor cares about me as a person.	1–7: Strongly disagree to Strongly agree	Mean of two support items.
Perceived emotional support	I feel understood by my instructor.	1–7: Strongly disagree to Strongly agree	Mean of two support items.

No item-level values were imputed. Composite summaries used only submitted survey items. Because the instruments were shortened or modified, all composites are exploratory descriptive indicators, not validated scales and not manipulation checks. The low-energy item was reverse-scored as specified in Supplementary Table S1.

Supplementary Appendix S2. Unit of Analysis, Data Preparation, and Four-Category Codebook

- Unit of analysis: One written response per one open-ended prompt. Each remaining response was coded into one mutually exclusive category.
- Cleaning: All non-student test entries, blank-code rows, blank or zero-width responses, placeholder-only responses, and lower-information duplicate attempts were removed. For duplicate attempts, the attempt with the greatest number of nonblank substantive responses was retained. Ties were resolved first by the greatest progress toward completion and then, if needed, by export order.
- Grade prefix text added by the platform prior to submission was removed before coding. The final analytical dataset consisted of 182 responses; 105 from Lesson A and 77 from Lesson B.
- Privacy: Participant-generated codes contained indirect information derived from a protected code-generation recipe and therefore were not treated as fully anonymous. No names or other direct identifiers were included in analytical or coding files. Randomized response IDs, randomized participant IDs, and neutral prompt IDs were used for coding and reporting. Participant-generated codes, identities, explicit lesson labels, original question numbers, and author ratings were removed from the independent coder's packet. Topic-specific biology content remained because it was required to interpret responses, so complete masking of lesson origin was impossible.

Supplementary Table S2. Four-Category Codebook

Category	Operational Definition	Decision Indicators
Developed	Relevant and meaningful; relates evidence or data to claims or explanations, describes causal or biological mechanisms, explains revisions to a model that have a meaningful biological or scientific purpose, and/or explains why or how an inquiry strategy was used.	Evidence-mechanism connection; causal explanation(s); meaningful revision(s); scientific purpose of the strategy.
Brief	Relevant and interpretable but limited; names a claim, pattern, factor, strategy, prediction, or revision with little explanation, evidence, mechanism, or scientific purpose.	Complete prompt-related idea without justification development.
Minimal	Some content from the prompt is present; however, the response is too fragmentary, vague, or unclear to permit interpretation of biology or scientific inquiry reasoning.	Single-word color response, unsupported fragment, or extremely incomplete answer.
Off-task	No substantive response to the prompt.	Unrelated joke, gibberish, emoji alone, refusal, or completely unrelated content.

Supplementary Appendix S3. Borderline Decision Rules

Supplementary Table S3. Borderline Decision Rules

Decision Boundary	Rule
Developed vs Brief	Use Developed only when the response provides a relationship between ideas (e.g., an evidence-mechanism connection), a meaningful revision, an evidence connection, and/or the scientific purpose of the strategy used. Simply naming a relevant idea or strategy is considered Brief.

Supplementary Table S3 (continued)

Decision Boundary	Rule
Brief vs Minimal	Use Brief when there is an interpretable claim, observation, factor, question, prediction, or strategy. Use Minimal when the content remains too vague or fragmentary to permit interpretation.
Minimal vs Off-task	All prompt-related content must be substantive. Use Off-task only if no prompt-related content can be recovered.
Multi-part prompts	Judge the best substantive reasoning present. A response need not address every sub-part of the original prompt to be considered Developed, but listing strategies without a rationale will typically remain Brief, and list fragments are considered Minimal.
Mixed relevant/off-task content	Code the most substantive relevant content. Use Off-task only if a joke or other unrelated material represents the entire response.
Factual errors	A factual error does not automatically lower category when developable reasoning is present.
Length	Length of response alone does not determine category. A shorter mechanism may be considered Developed while longer nonresponsive responses can still remain Minimal, Brief, or Off-task.

Supplementary Appendix S4. Coding Order, Blinding, and Provenance

1. The author coded all 182 responses before reviewing the independent coder's ratings.
2. The independent coder received a randomized packet containing response IDs, neutral prompt IDs, prompt text, and full retained response text. Participant-generated codes or identities, explicit lesson labels, original question numbers, and author-side ratings were removed. Topic-specific content remained where needed for interpretation; complete masking of lesson origin was therefore impossible.
3. The independent coder assigned one holistic category to each response without consultation with the author. Exact agreement and κ statistics were calculated before adjudication.
4. The coders reviewed the 55 disagreements using the frozen definitions and entered a separate consensus category. Neither original coder column was modified.

Supplementary Appendix S5. Interrater Agreement Before Adjudication

Category agreement was the primary measure of reliability; thus, the primary statistic used to adjust for chance was unweighted Cohen's κ . Linear- and quadratic-weighted κ values are presented as supplemental estimates because the four response-quality categories also have an ordinal structure.

Supplementary Table S5a. Agreement Statistics

Scope	N	Exact Agreements	Exact Agreement	Unweighted κ	Bootstrap 95% CI	Linear-weighted κ	Quadratic-weighted κ
Overall	182	127	69.8%	0.499	0.397–0.598	0.592	0.714
Lesson A	105	74	70.5%	0.468	0.330–0.604	0.553	0.674
Lesson B	77	53	68.8%	0.501	0.339–0.650	0.601	0.722

Supplementary Table S5b. Pre-Adjudication Agreement Matrix

Author-Side Set (Coder A) × Independent Coder (Coder B)	Developed	Brief	Minimal	Off-task	Row Total
Developed	71	21	0	0	92
Brief	7	51	5	1	64
Minimal	0	19	1	2	22
Off-task	0	0	0	4	4
Column Total	78	91	6	7	182

Of the 55 disagreements, 54 occurred across adjacent category boundaries: 28 Developed/Brief, 24 Brief/Minimal, and two Minimal/Off-task disagreements. The remaining disagreement was Brief/Off-task.

Supplementary Appendix S6. Prompt-Specific Worked Examples

De-identified samples were selected for each of the 21 open-ended prompts. Samples were purposively selected to demonstrate application of both the final category and decision rules; no separate numerical development scores or modal-category selection rules were utilized. These samples should not be considered prevalence estimates nor "the best" work of the students. All whitespace was normalized; the wording itself remained intact.

Supplementary Table S6. One De-Identified Worked Example for Each Open-Ended Prompt

Exempl e	Prompt ID	Prompt Text	De-identified Response	Final Category	Coding Rationale
A01	X16	Record your noticings and wonderings.	I noticed that the infrastructure is being built down and to the east. I question why the city isn't being built up, and what resources are comprised by the way the city is being built.	Developed	Developed because the response includes both an observation regarding direction of construction and a question which connects the observed pattern to potential resource consequences; it extends an isolated noticing.
A02	X14	Jot two questions down that you have about the effects of urbanization on non-human populations.	Will urbanization force organisms to move out of their natural habitats? Will urbanization affect the amount of resources available for non-human populations?	Brief	Brief because it presents two relevant questions regarding habitat displacement and resources, yet does not provide an explanation for an expected relationship or provide any evidence.
A03	X06	What nonhuman populations do you know that survive well in cities? Why do they do well? Jot down 1-2 sentences each.	Monkeys in Jaipur, India, survive well in the city because of the food surplus compared to forest. There is not only a factor of their swift movements that allow them to gather food, but the addition of monkeys association to religious gods in India.	Developed	Developed because it specifies food surpluses and human cultural interaction as two specific reasons that the population survives in an urban setting.
A04	X18	Why is making a prediction an important part of scientific inquiry? How does it help you think more clearly later on? Jot down your ideas.	Making a prediction helps you think about what you expect to happen using what you already know. It makes it easier to see if your ideas were correct or need to be changed after you observe the results.	Developed	Developed because it outlines how a prediction develops expectations which can subsequently be compared with subsequent observations and adjusted.

Supplementary Table S6 (continued)

Exempl e	Prompt ID	Prompt Text	De-identified Response	Final Category	Coding Rationale
A05	X01	What is urban? What is urbanization?	Urban is the idea of how people begin to form settlements and develop them into larger scales like cities. Urbanization is the development of these settlements to cities.	Brief	Brief because it defines both 'urban' and 'urbanization', but does not describe mechanisms, effects, or evidence.
A06	X12	What components in urban areas affect how well nonhuman populations survive? What types of interactions determine how well they survive?	Components in urban areas allow nonhumans to hide from predators and live off our scraps Interactions that keep them hidden while able to feed off our scraps	Developed	Developed because it connects shelter from predators and access to food scraps with interactions affecting survival.
A07	X10	How do you think increasing urbanization might affect nonhuman populations?	It can negatively impact many species, as urbanization can increase pollution and drive native species out of their habitats. However, it can also cause species well-adapted to urban areas to thrive.	Developed	Developed because it describes two different pathways: pollution and habitat displacement may adversely affect some species whereas urban adaptation may allow others to survive.
A08	X20	Choose a nonhuman population found in cities. Predict how the population has been OR will be affected by urbanization. Develop an initial model to explain how you think the population will survive OR has survived over time.	Monkeys in India at first would be generally shy of humans and forage in the forest living off natural diets, and they would establish social troops and trees and Rocky areas. But now there's an increased interaction between monkeys and humans. Monkeys have forage for food by snatching it from residents in tourists as well as scavenging from the garbage. They have become highly dependent on humid food, sources and the concentration of monkeys have led to a higher increase in disease.	Developed	Developed because the model relates increased human interaction, changed food sources, dependence upon humans, and disease risk to changes in the population over time.
A09	X13	Read the case study on Juncos How are urban and nonurban populations different? What cause might explain the difference(s) you observed? What questions do you have?	Urban juncos differ from non urban juncos, as urban juncos can be seen displaying signs of having bolder behavior, higher frequency songs, and shorter, thicker beaks. These differences may be caused by how they adapt to their new environment, as a higher frequency sound might be necessary to attract mates due to the loud sounds created by cities; bolder behavior may be developed by their interactions with humans, and the shorter, thicker beaks may be caused by their change in diets.	Developed	Developed because it connects observable junco characteristics with hypothesized mechanisms related to city noise levels, human interaction, and diets modified by cities.

Supplementary Table S6 (continued)

Exempl e	Prompt ID	Prompt Text	De-identified Response	Final Category	Coding Rationale
A10	X04	Answer in 1-2 sentence: How is the case study you read similar or different to your initial model?	It's similar in that it focuses on 1 aspect. I wanted to focus on living space, the study focused on boldness.	Brief	Brief because it lists a single similarity and a single difference between the case and the initial model and lacks description of the significance of those differences.
A11	X21	On your own Revise your model. What details/causes/mechanisms can you add, if any?	At first, pigeons served as messengers for humans in towns and cities. Now as urbanization has spread, the pigeons rely on scraps from the streets instead of the usual food they would have been given by their owners. Their flight initiation distance would also differ as they would be more exposed to humans over time, becoming bolder in their behavior and approaching more unfamiliar objects as a resort for scavenging for themselves.	Developed	Developed because the revision incorporates mechanisms related to availability of food, repeated exposure to humans, decreased flight distances, and increased boldness.
A12	X05	Choose one strategy we used today (modeling, revising, or using evidence). Explain why scientists use this strategy and how you used it today.	Modeling is used by scientists because it helps them visualize their understandings and what they could improve on in their initial analysis. I used modeling today to represent the changes in raccoon and junco behaviors as they adapt to their new environment.	Developed	Developed because it describes the scientific objective of modeling as well as the way modeling was employed to represent behavioral changes.
B01	X09	Jot down anything you noticed about these graphs:	rate of new cancers is more primarily more dense on the east coast.	Brief	Brief because it identifies a relevant geographically-dependent pattern in the graph but does not identify supportive data or explain the identified pattern.
B02	X03	What is one question that you have whose answer will help us make progress on our lesson question: Who gets cancer and why?	What kind of cancer are we talking about?	Brief	Brief because it provides a relevant clarifying question but no claim, evidence, or explanatory reasoning.
B03	X08	What did you notice about the correlation between height and cancer risk?	The taller a person the more likely they are to get cancer	Brief	Brief because it states a directional relationship between height and cancer risk without evidence or a biological explanation.
B04	X17	What kind of claims can you make? What evidence do you have to make those claims?	It seems like there is no correlation the evidence is the graph for males from 160 cm to 175 cm	Developed	Developed because it pairs a tentative claim with specific graph data related to stated male height range as well as crossing the Brief-to-Developed boundary via an explicit claim-evidence link.

Supplementary Table S6 (continued)

Exempl e	Prompt ID	Prompt Text	De-identified Response	Final Category	Coding Rationale
B05	X15	Why do scientists use animal models, such as mice? Based on the claims you made about the relationships between height and age and incidence of cancer in humans, what kinds of mammals would you predict have the highest incidence of cancer?	Scientists use mice because they are mammals with a high genetic similiarity to humans, the oldest mammals	Brief	Brief because it provides a relevant reason for using mice and a directional relationship between age and cancer risk without any supporting evidence or biologic explanation.
B06	X11	What predictions did scientists make about cancer rates in nonhuman animals? What did they observe? What mechanisms did they propose that might explain their observations? Here is the reading: [URL] Incidence of Cancer	Scientists expected large and longevous animals to have higher incidence of cancer. Scientists observed data collected from large and long-lived whales and elephants and found that they rarely develop cancer. Scientists proposed that extra copies of a gene called p53 found in large, long-lived animals which is involved in protecting mammals from developing cancer might explain their rare incidence of developing cancer.	Developed	Developed because it differentiates among the original prediction, observed whale and elephant pattern, and proposed protective mechanism based on p53.
B07	X19	Develop a model using pictures, symbols, and/or words to explain: Who gets cancer and why? Include the following: Components Interactions Mechanisms	In terms of who gets cancer and why, the ones who are most likely to get cancer are adult humans which are generally caused by harmful daily habits like smoking and genetic mutations that code to create the spread of cancerous cells. The ones who don't cancer as often are those who have a decent if not abundant amount of the p53 gene, allowing them a sort of protection against metastasis.	Developed	Developed because the model relates smoking and genetic mutations to cancerous-cell spread while also contrasting this with a proposed protective role for p53.
B08	X02	What does your model explain? What additional information do you need to explain Who gets cancer and why?	it explains the correlation of recieving cancer due to age. i would also need to test other factors (maybe on the type of cancer)	Brief	Brief because it identifies age as a factor in addition to recognizing that other factors must be tested but does not explain any mechanisms.
B09	X07	Choose one strategy we used today (modeling, revising, or using evidence). Explain why scientists use this strategy and how you used it today.	Scientists use modeling to simplify ideas and understand patterns. It helps them explain how different factors connect and make predictions. Today, I used modeling by explaining how age, genetics, and environment interact to cause cancer through DNA mutations.	Developed	Developed because it explains why scientists use models and applies the strategy to interactions among age, genetics, environment, and DNA mutations.

Supplementary Appendix S7. Final Adjudicated Counts by Prompt

Supplementary Table S7. Final Adjudicated Response Counts by Prompt

Lesson	Question	Prompt ID	Responses	Developed	Brief	Minimal	Off-task
A	1	X16	9	4	4	0	1
A	2	X14	9	0	8	0	1
A	3	X06	9	8	1	0	0
A	4	X18	9	7	2	0	0
A	5	X01	9	1	8	0	0
A	6	X12	9	5	4	0	0
A	7	X10	9	5	4	0	0
A	8	X20	9	6	3	0	0
A	9	X13	9	6	2	1	0
A	10	X04	8	3	4	1	0
A	11	X21	8	3	3	2	0
A	12	X05	8	6	2	0	0
B	1	X09	10	3	5	2	0
B	2	X03	10	0	10	0	0
B	3	X08	9	2	6	0	1
B	4	X17	9	5	4	0	0
B	5	X15	9	3	6	0	0
B	6	X11	9	5	2	0	2
B	7	X19	7	3	3	0	1
B	8	X02	7	0	6	0	1
B	9	X07	7	4	3	0	0

Supplementary Appendix S8. Final Adjudicated Counts by Randomized Participant Identifier

Participant ID numbers listed below are randomized participant IDs generated for analytical-reporting purposes. These numbers are anonymized identifiers generated by us for reporting and cannot be used as a lookup for identifying actual names. Table S8 lists the 13 participants who submitted at least one retained coded response; the two other valid pseudonymous participants did not submit any retained written responses and will therefore not appear in the response-count table.

Supplementary Table S8. Final Adjudicated Response Counts by Randomized Participant Identifier

Random Participant ID	Lesson A Responses	Lesson B Responses	Total	Developed	Brief	Minimal	Off-task
P01	12	0	12	8	3	1	0
P02	12	9	21	17	4	0	0
P03	0	2	2	1	1	0	0
P04	9	9	18	9	8	0	1
P05	12	9	21	7	14	0	0
P06	12	9	21	2	12	3	4
P07	12	9	21	6	14	1	0
P08	0	6	6	0	5	0	1
P09	12	0	12	6	5	1	0
P10	12	9	21	14	7	0	0
P11	0	9	9	0	9	0	0
P12	0	6	6	3	3	0	0
P13	12	0	12	6	5	0	1