Supporting Information for

Conditions for Effective Learning Without Upfront Instruction: How Practice with Feedback Supports Memory, Generalization, and Metacognition

Table of Contents

Additional supporting information can be found at
https://osf.io/kruzw/?view_only=bbef5ded9a834e089b54bd23f9ac342a

Study 1

Descriptive Statistics

Table S1

Zero-Order Correlations and Descriptive Statistics, Study 1

Measure	1	2	3	4	5	6	7	8	9	10	11	12
1. Baseline Conf.												
2. Baseline Int.	.69											
3. Pretest Score	.04	-.11										
4. Gen. Score	.34	.18	.36									
5. Instruction	.15	.16	-.15	-.02								
6. JOL	.50	.51	.00	.27	.11							
7. Regression Conf.	.56	.51	.09	.35	.07	.60						
8. Expectancy	.57	.51	.03	.38	.07	.72	.67					
9. SI in Session	.47	.52	.02	.21	.17	.68	.60	.61				
10. Regression Int.	.56	.70	-.04	.16	.17	.60	.63	.60	.78			
11. UV	.47	.53	-.03	.19	.12	.57	.58	.53	.60	.71		
12. Distraction	-.19	-.18	-.17	-.20	-.05	-.25	-.27	-.23	-.44	-.29	-.18	
M	3.28	3.29	0.58	0.33	8.23	4.62	3.67	3.14	4.68	3.55	3.77	2.09
SD	0.93	0.97	0.37	0.29	4.31	1.04	0.94	1.05	0.98	1.19	0.84	1.13

Note. Conf. = Confidence, Int. = Interest, Mem. = Memory, Gen. = Generalization, JOL = Judgment of Learning, SI = Situational Interest, UV = Utility Value. All correlations $| r | \geq .12$ are significant at $p < .05$.

Table S2

Study 1 Means and Standard Deviations in the Lecture (N = 100), Practice with Correct Response Feedback (N = 100), and Practice with Explanatory Feedback (N = 97) conditions.

	Lecture		Practice + CR Feedback		Practice + Explanatory Feedback	
Outcome	Mean	SD	Mean	SD	Mean	SD
Memory Score (Pct.)	0.46	0.38	0.65	0.34	0.64	0.35
Generalization Score (Pct.)	0.39	0.31	0.27	0.26	0.33	0.29
Instruction Time	9.90	3.33	7.25	4.69	7.51	4.34
Judgment of Learning	4.69	0.90	4.35	1.17	4.84	1.00
Regression Confidence	3.73	0.93	3.56	0.97	3.71	0.91
Expectancy for Success	3.23	1.00	3.00	1.04	3.19	1.12
S.I. in Session	4.71	0.89	4.60	0.99	4.74	1.05
Interest in Regression	3.57	1.16	3.47	1.11	3.64	1.31
UV for Regression	3.81	0.85	3.67	0.80	3.82	0.88
Distraction	2.12	1.19	2.21	1.10	1.92	1.07

Note. Pct. = Percent. SI = Situational Interest, UV = Utility Value. CR = Correct Response.

Linear Models for Primary Outcomes

Below we report regression results for all measured variables, including two exploratory not discussed in the main text: utility value for regression and self-reported distraction. There are no significant effects of condition on either manipulation.

Table S3

Memory and Generalization Performance, Study 1

	Memory Percent				Generalization Percent			
Term	b	se	t	p	b	se	t	p
Intercept	0.64	0.04	17.39	.000	0.32	0.03	11.79	.000
Lecture (vs. Exp. Feedback)	-0.17	0.05	-3.39	.001	0.06	0.04	1.51	.131
CR Feedback (vs. Exp. Feedback)	0.01	0.05	0.15	.882	-0.04	0.04	-1.17	.245
Confidence (Standardized)	0.01	0.04	0.29	.768	0.12	0.03	4.53	.000
Lecture x Confidence	0.02	0.05	0.32	.748	0.01	0.04	0.22	.829
CR Feedback x Confidence	-0.04	0.05	-0.81	.421	-0.08	0.04	-2.22	.027

Note. Exp. = Explanatory, CR = Correct Response

Table S4

Instruction Time (in Minutes), Study 1

Term	b	se	t	p
Intercept	7.50	0.42	17.93	.000
Lecture (vs. Exp. Feedback)	2.35	0.59	4.00	.000
CR Feedback (vs. Exp. Feedback)	-0.11	0.59	-0.19	.848
Confidence (Std.)	0.12	0.42	0.28	.782
Lecture x Conf.	0.43	0.60	0.72	.474
CR Feedback x Conf.	0.96	0.58	1.65	.101

Table S5

Judgment of Learning, Regression Confidence, and Expectancies for Success, Study 1

	Judgement of Learning				Regression Confidence			
Term	*b*	*se*	*t*	*p*	*b*	*se*	*t*	*p*
Intercept	4.81	0.09	53.23	.000	3.68	0.08	46.47	.000
Lecture (vs. Exp. Feedback)	-0.16	0.13	-1.24	.217	0.00	0.11	0.04	.967
CR Feedback (vs. Exp. Feedback)	-0.38	0.13	-3.00	.003	-0.06	0.11	-0.51	.609
Confidence (Std.)	0.46	0.09	5.09	.000	0.49	0.08	6.19	.000
Lecture x Conf.	-0.03	0.13	-0.24	.809	0.06	0.11	0.56	.576
CR Feedback x Conf.	0.18	0.13	1.45	.147	0.04	0.11	0.39	.696

	Expectancies for Success on Test			
	b	*se*	*t*	*p*
Intercept	3.16	0.09	35.78	.000
Lecture (vs. Exp. Feedback)	0.02	0.12	0.17	.862
CR Feedback (vs. Exp. Feedback)	-0.08	0.12	-0.68	.498
Confidence (Std.)	0.61	0.09	6.96	.000
Lecture x Conf.	-0.03	0.13	-0.20	.840
CR Feedback x Conf.	0.00	0.12	-0.03	.977

Table S6

Situational Interest, Interest in Regression, Utility Value for Regression, Self-Reported Distraction, Study 1

	SI in Session				Interest in Regression			
Term	*b*	*se*	*t*	*p*	*b*	*se*	*t*	*p*
Intercept	4.71	0.09	53.77	.000	3.60	0.10	35.79	.000
Lecture (vs. Exp. Feedback)	-0.03	0.12	-0.24	.812	-0.09	0.14	-0.61	.545
CR Feedback (vs. Exp. Feedback)	-0.06	0.12	-0.45	.655	-0.06	0.14	-0.42	.674
Confidence (Std.)	0.54	0.09	6.20	.000	0.79	0.10	7.93	.000
Lecture x Conf.	-0.21	0.12	-1.69	.092	-0.15	0.14	-1.01	.311
CR Feedback x Conf.	-0.05	0.12	-0.37	.711	-0.21	0.14	-1.49	.136

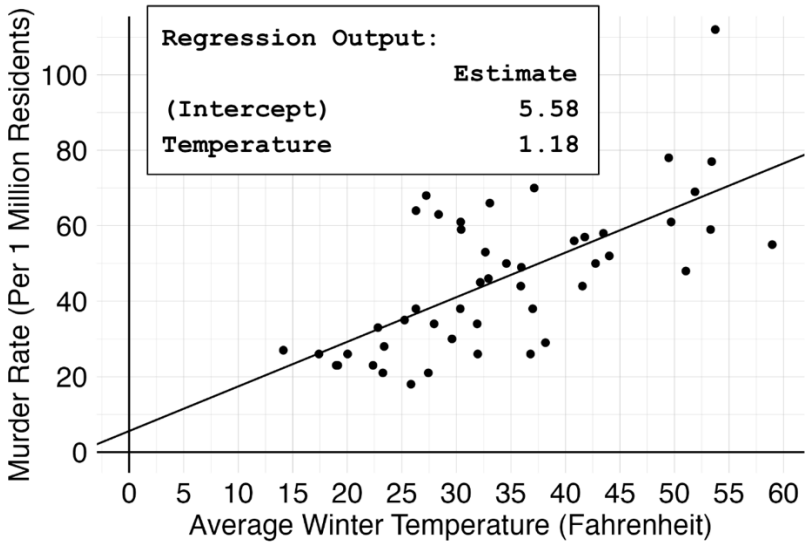
	UV for Regression				Self-Reported Distraction			
	b	*se*	*t*	*p*	*b*	*se*	*t*	*p*
Intercept	3.80	0.08	50.00	.000	1.94	0.11	17.28	0.00
Lecture (vs. Exp. Feedback)	-0.02	0.11	-0.23	.816	0.21	0.16	1.35	0.18
CR Feedback (vs. Exp. Feedback)	-0.08	0.11	-0.79	.432	0.27	0.16	1.72	0.09
Confidence (Std.)	0.46	0.08	6.08	.000	-0.26	0.11	-2.36	0.02
Lecture x Conf.	-0.12	0.11	-1.12	.265	-0.08	0.16	-0.48	0.63
CR Feedback x Conf.	-0.08	0.11	-0.73	.465	0.24	0.16	1.54	0.13

Posttest Materials and Scoring Rubric

All questions and scoring criteria for the Study 1 posttest are provided below. The posttest was scored by the lead author of the manuscript using these criteria. All responses earned 1 point if criteria were met and zero points if they were not.

Part 1: Review [Memory Questions]

In the scatterplot below, each point represents a U.S. state. The plot shows the relationship between each state's average temperature and its murder rate.



1A: What does it mean that the intercept is 5.58?

- **For a state with average winter temperatures of 0 degrees, the predicted murder rate is 5.58.**
- The murder rate increases by 5.58 units for every 1 degree increase in average temperature.
- 5.58 is where the murder rate starts.
- 5.58 is the lowest murder rate for any state.

1B: What does it mean that the slope for temperature is 1.18?

- **For every 1 degree increase in average temperature, the murder rate increases by 1.18 units.**
- For a state with average winter temperatures of 0 degrees, the predicted murder rate is 1.18.
- 1.18 is the average murder rate.
- For every 10 degree increase in average temperature, the murder rate increases by 1.18 units.

1C: 1C: What is the predicted murder rate for a state with an average winter temperature of 20 degrees?

- **29.18**
- 20.00
- 28.76
- 5.58

Part 2 [Generalization Questions]

For the next 3 questions: A regression between high school GPA and college GPA resulted in the following equation:

$$\text{CollegeGPA} = 1.097 + 0.675 * \text{HighSchoolGPA}$$

2A: Using this equation, what is the predicted college GPA for a student with a high school GPA of 3.25? Round your answer to two decimal places.

- **Scoring Rule:** Any answer that rounded to 3.3 should be counted as correct.

2B: What is the meaning of the slope in this equation?

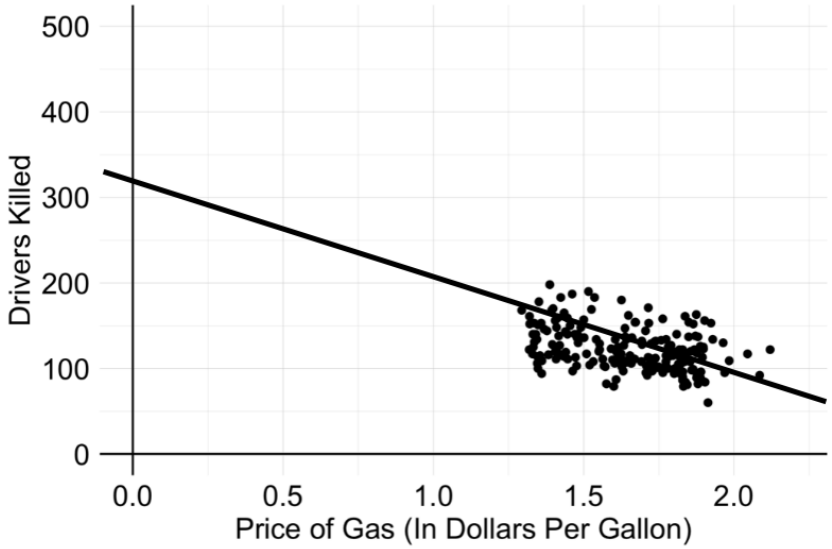
- **Scoring Rule:** A correct answer should communicate that a one-unit change in the predictor variable (High School GPA) is associated with a .675-unit increase in the outcome variable (College GPA). If the student's answer doesn't specify or imply a one-unit change in high school GPA, it is wrong. A correct answer should mention high school and college GPA, but it's okay if the student is vague with these labels. It's also okay if the student rounds .675 to any number between .6 and .7, and it's okay if they omit the number .675 entirely but say something like "the slope is the change in College GPA for each 1-point increase in high school GPA." Here are two examples of correct answers: "A one unit change in high school GPA is associated with a .675 unit increase in college GPA" or "for every increase in high school GPA, college GPA goes up by .675."2C: What is the meaning of the intercept in this equation?

2C: What is the meaning of the intercept in this equation?

- **Scoring Rule:** A correct answer should communicate that for a student with a high school GPA of 0, their predicted college GPA is 1.097. It's okay if the student rounds 1.097 to any number between 1 and 1.1. Finally, it's okay if they omit the specific number for the intercept entirely but say something like "it shows the college GPA for someone who had a high school GPA of 0." Here are two examples of correct answers: "If high school GPA is 0, college GPA is 1.1" or "A student who didn't go to high school would have a college GPA of 1.097."

Part 3 [Generalization Questions]

For the next 5 questions. The scatterplot below comes from a study examining factors that predict traffic fatalities. Each point represents a geographic region. The graph below shows the relationship between the price of gas (in dollars per gallon) in that region and the number of drivers killed in a given month.



3A. Estimate the slope and intercept of the regression line on the graph:

- **Slope Scoring Rule:** Any number between -85 and -145 was counted as correct
- **Intercept Scoring Rule:** Any number between 300 and less than 340 was counted as correct.

3B. What is the meaning of the slope?

- **Scoring Rule:** A correct answer should mention drivers killed and gas prices, but it's okay if the student is vague with these labels or uses different words. It is also okay if the student uses any number between -85 and -145 for the slope. Also, if the student estimated the slope incorrectly in the first step of the problem, we don't want to penalize them twice for the same mistake so it's okay if they reuse the same incorrect number from their first answer. If the student's answer doesn't specify or imply a one-unit change in gas prices, the answer is wrong. However, it is okay if they omit the specific number for the slope entirely but say something like "the slope is the change in deaths for each dollar increase in gas prices." Here are two examples of correct answers: "A one unit change in gas prices is associated with -90 unit change in drivers killed." or "for every one dollar increase in gas prices, traffic deaths drop by 120."

3C: What is the meaning of the intercept?

- **Scoring Rule:** A correct answer should communicate that when gas prices are zero, there would be 320 drivers killed. It's okay if the student uses any number between 305 and 335 for the intercept. Also, if the student estimated the intercept incorrectly in the first step of the problem, we don't want to penalize them twice for the same mistake so it's okay if they reuse the same incorrect number from their first answer. Finally, it's okay if they omit the specific number for the intercept entirely but say something like "it shows the number of drivers killed if gas were free." Here are two examples of correct answers: "When gas costs $0, there are 320 traffic deaths" or "330 drivers would die if gas became free."

3D: Write a linear equation representing the regression line on the graph, using the form y = b0 + b1x

- **Scoring Rule:** A correct answer could substitute the students previous answer for the intercept in for "b0" and their previous answer for the slope in for "b1." A correct answer could also use any value between 305 and 335 for the intercept, and between -85 and -145 for the slope. It's okay if the student leaves "y" and "x" unchanged in their answer, or if they substitute something involving "drivers killed" for "y" and something involving "gas price" for "x." For example, if a student previously estimated that the intercept was 310 and the slope was -100, any of the following answers would be correct: "y = 310 – 100 * x" or "deaths = 310 – 100 * gas" or even "y = 310 – 100 * gas price." However, it's incorrect if a student substitutes something like "gas" in for "y" or something like "drivers killed" for "x."

3E: Using the linear equation that you generated, estimate the number of traffic fatalities we might observe if the price of gas dropped to $1.20 per gallon. Round your estimate to the nearest whole number, and enter it in the box below:

- **Scoring Rule:** A correct response should use the student's previous answers for the slope and intercept to make the prediction, using the following formula. Fatalities = intercept + slope * 1.20. Any answer that rounds to the nearest 10 is be counted as correct.

Study 2

Descriptive Statistics

Table S7

Zero-Order Correlations and Descriptive Statistics, Study 2

Measure	1	2	3	4	5	6	7	8	9	10	11	12
1. Baseline Conf.												
2. Baseline Int.	.59											
3. Pretest Score	.04	-.11										
4. Mem. Score (Pct.)	.06	.01	.32									
5. Gen. Score (Pct.)	.04	-.09	.30	.66								
6. Instruction Time	.10	.14	-.11	.00	.07							
7. JOL	.49	.54	-.02	.03	.06	.22						
8. Overconfidence	.32	.35	-.21	-.62	-.63	.15	.42					
9. S.I. in Session	.40	.61	-.01	.08	.08	.22	.72	.31				
10. Regression Int.	.45	.70	-.08	.03	-.02	.25	.65	.34	.78			
11. UV	.41	.58	-.04	.03	-.02	.24	.59	.31	.71	.82		
12. Distraction	.02	-.04	-.20	-.12	-.07	.08	-.11	.08	-.24	-.11	-.14	
M	3.07	3.27	.39	0.44	0.35	6.61	4.38	0.16	4.51	3.54	3.59	2.06
SD	0.98	0.97	0.31	0.36	0.35	3.30	1.36	0.36	1.15	1.25	1.05	1.20

Note. Conf. = Confidence, Int. = Interest, Mem. = Memory, Gen. = Generalization, JOL = Judgment of Learning, SI = Situational Interest, UV = Utility Value. All correlations $|r| \geq .12$ are significant at $p < .05$.

Table S8

Study 2 Means and Standard Deviations in the Lecture (N = 100), Practice with Correct Response Feedback (N = 100), and Practice with Explanatory Feedback (N = 100) conditions.

	Lecture		Practice + CR Feedback		Practice + Explanatory Feedback	
Outcome	Mean	SD	Mean	SD	Mean	SD
Memory Score (Pct.)	0.36	0.36	0.45	0.37	0.50	0.35
Generalization Score (Pct.)	0.36	0.35	0.30	0.35	0.39	0.36
Instruction Time	8.79	1.88	5.09	3.02	6.03	3.55
Judgment of Learning	4.59	1.13	4.01	1.53	4.53	1.31
Overconfidence	0.26	0.36	0.10	0.35	0.13	0.37
S.I. in Session	4.66	1.04	4.29	1.23	4.59	1.16
Interest in Regression	3.70	1.27	3.41	1.30	3.51	1.18
UV for Regression	3.76	1.06	3.50	1.02	3.53	1.07
Distraction	2.06	1.14	2.08	1.17	2.05	1.29

Linear Models for Primary Outcomes

Table S9

Memory and Generalization Performance, Study 2

	Memory Percent				Generalization Percent			
Term	b	se	t	p	b	se	t	p
Intercept	0.50	0.03	14.72	.000	0.39	0.03	11.74	.000
Lecture (vs. Exp. Feedback)	-0.13	0.05	-2.64	.009	-0.02	0.05	-0.51	.609
CR Feedback (vs. Exp. Feedback)	-0.05	0.05	-0.99	.325	-0.09	0.05	-1.97	.050
Pretest Score (Standardized)	0.11	0.03	3.20	.002	0.11	0.03	3.35	.001
Lecture x Pretest	0.04	0.05	0.72	.475	0.05	0.05	1.06	.290
CR Feedback x Pretest	-0.01	0.05	-0.27	.784	-0.06	0.05	-1.28	.202

Note. Exp. = Explanatory, CR = Correct Response

Table S10

Instruction Time (in Minutes), Study 2

Term	b	se	t	p
Intercept	6.00	0.29	20.75	.000
Lecture (vs. Exp. Feedback)	2.77	0.42	6.67	.000
CR Feedback (vs. Exp. Feedback)	-0.89	0.41	-2.16	.031
Pretest Score (Std.)	0.28	0.29	0.94	.346
Lecture x Pretest	-0.05	0.44	-0.12	.905
CR Feedback x Pretest	-0.13	0.40	-0.34	.737

Table S11

Efficiency for Memory and Generalization, Study 2

	Efficiency: Memory					Efficiency: Generalization				
Term	b	se	t	df	p	b	se	t	df	p
Intercept	6.93	0.76	9.10	291.99	.000	4.27	0.72	5.94	291.99	.000
Lecture (vs. Exp. Feedback)	-1.63	1.08	-1.50	291.99	.134	-2.16	1.02	-2.11	291.99	.036
CR Feedback (vs. Exp. Feedback)	-4.84	1.14	-4.22	265.80	.000	-2.24	1.09	-2.06	262.87	.041
Pretest Score (Standardized)	0.30	0.79	0.38	291.99	.704	0.52	0.74	0.70	291.99	.486
Lecture x Pretest	0.99	1.07	0.92	291.99	.358	0.00	1.01	0.00	291.99	.998
CR Feedback x Pretest	1.45	1.19	1.22	264.24	.224	1.44	1.13	1.28	261.16	.200

Table S12

Judgment of Learning and Overconfidence, Study 2

	Judgment of Learning				Overconfidence			
Term	b	se	t	p	b	se	t	p
Intercept	4.49	0.12	38.84	.000	0.12	0.03	3.56	.000
Lecture (vs. Exp. Feedback)	0.04	0.17	0.24	.810	0.13	0.05	2.69	.008
CR Feedback (vs. Exp. Feedback)	-0.37	0.17	-2.24	.026	-0.01	0.05	-0.11	.912
Pretest Score (Std.)	0.49	0.12	4.17	.000	0.12	0.03	3.57	.000
Lecture x Pretest	0.26	0.18	1.49	.137	0.03	0.05	0.49	.627
CR Feedback x Pretest	0.22	0.16	1.39	.165	-0.04	0.05	-0.76	.449

Table S13

Situational Interest, Interest in Regression, Utility Value for Regression, Self-Reported Distraction, Study 1

	SI in Session				Utility Value for Regression			
Term	b	se	t	p	b	se	t	p
Intercept	4.57	0.10	43.92	.000	3.47	0.11	31.36	.000
Lecture (vs. Exp. Feedback)	0.05	0.15	0.33	.743	0.18	0.16	1.12	.262
CR Feedback (vs. Exp. Feedback)	-0.20	0.15	-1.36	.174	0.03	0.16	0.16	.869
Pretest Score (Standardized)	0.29	0.11	2.72	.007	0.45	0.11	4.03	.000
Lecture x Pretest	0.32	0.16	2.03	.043	0.21	0.17	1.24	.218
CR Feedback x Pretest	0.18	0.14	1.28	.201	0.14	0.15	0.90	.369

	Self-Reported Distraction			
	b	se	t	p
Intercept	3.51	0.09	37.06	.000
Lecture (vs. Exp. Feedback)	0.20	0.14	1.51	.133
CR Feedback (vs. Exp. Feedback)	0.05	0.14	0.39	.698
Pretest Score (Standardized)	0.32	0.10	3.29	.001
Lecture x Pretest	0.31	0.15	2.15	.032
CR Feedback x Pretest	0.07	0.13	0.51	.609

Preregistered Analyses of Instruction Time

Participants in the explanatory feedback condition completed the learning session in an average of 6.0 minutes, compared to 8.8 minutes in the lecture condition, $t(294) = 6.67$, $p < .001$—a 35% time savings. Explanatory feedback also took slightly longer than correct-response feedback (M = 6.0 vs. 5.1 minutes), $t(294) = -2.16$, $p = .031$, driven by participants spending more time reviewing the longer, explanation-based feedback (1.3 minutes vs. 0.5 minutes, on average). No condition × baseline confidence interactions were significant ($ps \geq .737$). On average, practice with correct response feedback was 3.7 minutes faster than lecture, $t(294) = -8.23$, $p < .001$.

Details, Calculating Efficiency

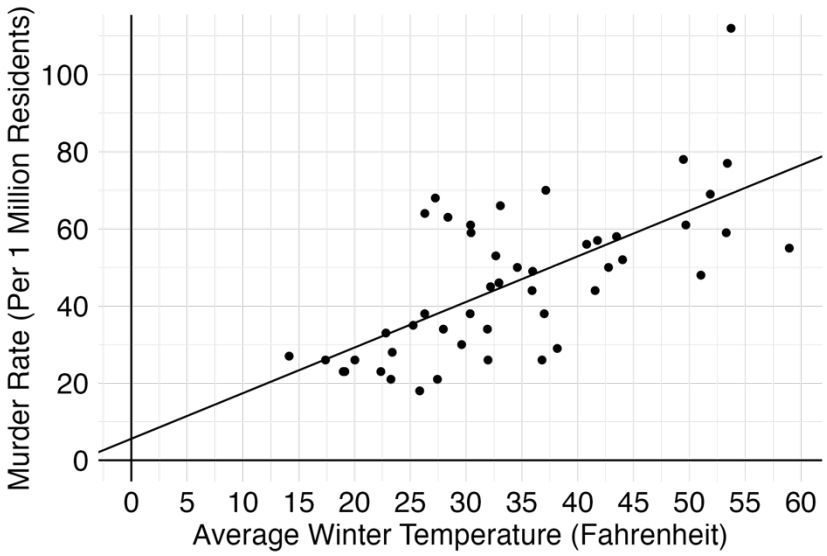
To estimate initial knowledge for participants in the practice conditions, we used participants' performance on the practice problems. This estimate was not available for participants in the lecture conditions because students did not complete a practice problems. However, because students were randomly assigned to condition, we could confidently assume that data about initial knowledge were missing completely at random and use a multiple imputation procedure to repeatedly sample plausible values from the distribution of estimated values in the practice conditions. We then computed efficiency scores for participants in all imputed datasets and analyzed them separately, pooling regression estimates with Rubin's rules (Rubin, 1987) to yield an unbiased estimate of efficiency while reflecting the uncertainty in many participants' true initial knowledge. The imputation procedure was conducted with the "mice" package in R (vanBurren & Groothuis-Oudshoorn, 2011). To obtain predicted values for Figure 7, we followed the "Predict then Combine" recommendations of Miles (2016).

Posttest Materials and Scoring Rubric

All questions and scoring criteria for the Study 2 posttest are provided below. The posttest was scored by the lead author of the manuscript using these criteria. All responses earned 1 point if criteria were met and zero points if they were not.

Memory Questions

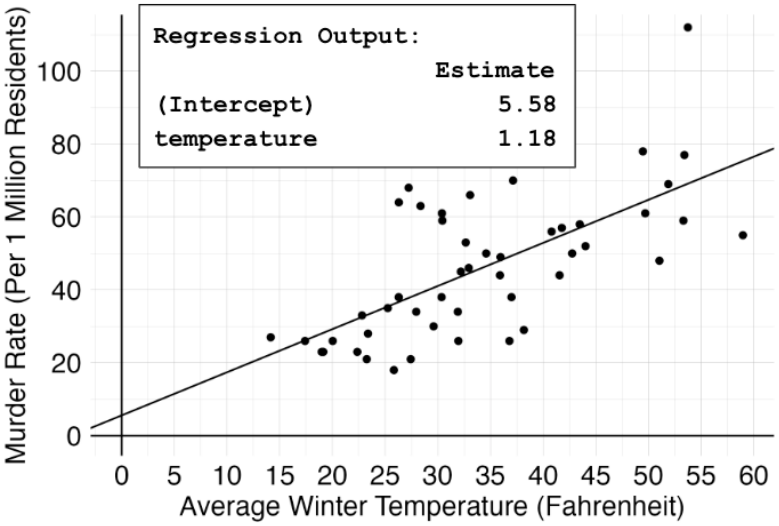
In the scatterplot below, each point represents a U.S. state. The plot shows the relationship between each state's average temperature and its murder rate.



1A: A regression line, determined by a computer, is shown on the plot. What is the approximate intercept of the line?

- **Scoring Rule:** Any answer between 5 and 7 (inclusive) should be counted as correct.

In the scatterplot below, each point represents a U.S. state. The plot shows the relationship between each state's average temperature and its murder rate.



The exact, estimated slope and intercept for the regression line are now displayed on the figure.

1B: What does it mean that the intercept is 5.58?

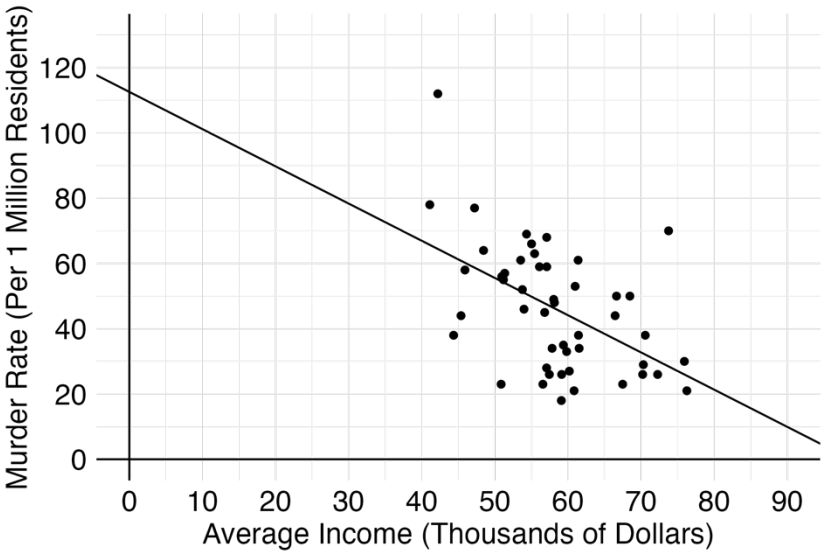
- **Scoring Rule:** A correct answer should communicate for a state where average winter temperature is zero, there are an estimated 5.58 murders per million residents. It's okay if the student rounds this to 5 or 6. It's okay if the student only refers to "temperature" (without mentioning winter) or "murders" (instead of murder rate). Finally, it's okay if they omit the specific number for the intercept entirely but say something like "it shows the number of murders if winter temperature were 0." Here is an example of a correct answer: "When temperature is 0, murder rate would be 5.58."

1C: What does it mean that the slope for temperature is 1.18?

- **Scoring Rule:** A correct answer should mention murder rates and winter temperature, but it's okay if the student is vague with these labels or uses different words. If the student's answer doesn't specify or imply a one-unit change in gas prices, the answer is wrong. However, it is okay if they omit the specific number for the slope entirely but say something like "the slope is the change in murder rate for each degree change in temperature." Here are two examples of correct answers: "A one unit change in temperature is associated with 5.58 fewer murders per million." or "for every one degree increase in temperature, murder rates drop by 5.58."

Generalization Questions

In the scatterplot below, each point represents a U.S. state. The plot shows the relationship between each state's average family income (in thousands of dollars) and its murder rate.



A regression line, determined by a computer, is shown on the plot. What is the approximate intercept of the line?

- **Scoring Rule:** A correct answer gives a number between 110 and 115 (inclusive)

In the scatterplot below, each point represents a U.S. state. The plot shows the relationship between each state's average family income (in thousands of dollars) and its murder rate.

The exact, estimated slope and intercept for the regression line are now displayed on the figure.

What does it mean that the intercept is 112.6?

- **Scoring Rule:** A correct answer should communicate that the intercept tells us the predicted murders (or murder rate) when income equals zero. They don't need to say

112.6 in their response (and it's okay if they round the number to 113, or say something like "about 110"), but they do need to say something about income equaling zero. Here is an example of a correct answer: "When average income is 0, murder rate would be about 113."

What does it mean that the slope for income is -1.1?

- **Scoring Rule:** A correct answer should communicate that the slope tells us the change in murders (or the murder rate) when income increases by 1 (or 1 thousand dollars). They don't need to say 112.6 in their response (and it's okay if they round the number to 113 or say something like "about 110"), but they do need to say something about a 1-unit or 1-thousand-dollar change in income. Here are two examples of correct answers: "A one unit decrease in income is associated with about 1.1 fewer murders per million." or "for every thousand dollars of income, murder rates drop by 1.1."

LLM Feedback Prompts for Practice Questions

Practice Question 1:

```
[
    {
     "role": "system",
     "content": "A student was shown a scatterplot with the description, “In the scatterplot below, each point represents a U.S. state.” They were then given the following prompt: “The plot shows the relationship between each state’s (winter) average temperature and its murder rate. A regression line, determined by a computer, is shown on the plot. What is the approximate intercept of the line?” The intercept was approximately 6, but you can count anything between 5 and 7 as correct. Give them feedback about the correctness of their answer. Your feedback should contain only the following: (1) The word “Correct!” or “Incorrect.” (2) if the answer is wrong, explain exactly why. (3) if the answer is wrong, then say “Corrected version:” and edit their exact wording with the smallest possible changes to make their response fully correct. Address the student directly."
    },
    {
     "role": "user",
     "content": "Here is the student's response: [RESPONSE HERE]"
    }
 ]
```

Practice Question 2:

```
[
    {
     "role": "system",
     "content": "A student was shown a scatterplot with the description, “In the scatterplot below, each point represents a U.S. state. “The plot shows the relationship between each state’s average (winter) temperature and its murder rate. A regression line, determined by a computer, is shown on the plot.”  They were told that the slope of the regression line was 1.18, and that the intercept was 5.58. They were then asked: “What does it mean that the intercept is 5.58?” Give them feedback about the correctness of their answer. Your feedback should contain only the following: (1) The word “Correct!” or “Incorrect.” (2) if the answer is wrong, explain exactly why. (3) if the answer is wrong, then say “Corrected version:” and edit their exact wording with the smallest possible changes to make their response fully correct. A correct answer should communicate that the intercept tells us the predicted murders (or murder rate) when temperature equals zero. They don’t need to say 5.58 in their response (and it’s okay if they round the number to 6, or say something like “between 5 and 6”), but they do need to say something about temperature equaling zero. Address the student directly."
    },
    {
     "role": "user",
     "content": "Here is the student's response: [RESPONSE HERE]"
```

}]

Practice Question 3:

[
{
"role": "system",
"content": "A student was shown a scatterplot with the description, “In the scatterplot below, each point represents a U.S. state. “The plot shows the relationship between each state’s average (winter) temperature and its murder rate. A regression line, determined by a computer, is shown on the plot.” They were told that the slope of the regression line was 1.18, and that the intercept was 5.58. They were then asked: “What does it mean that the slope for temperature is 1.18?” Give them feedback about the correctness of their answer. Your feedback should contain only the following: (1) The word “Correct!” or “Incorrect.” (2) if the answer is wrong, explain exactly why. (3) if the answer is wrong, then say “Corrected version:” and edit their exact wording with the smallest possible changes to make their response fully correct. A correct answer should communicate that the slope tells us the change in murders (or the murder rate) when temperature increases by 1. They don’t need to say 1.18 in their response (and it’s okay if they round the number to 1.2, or say something like “between 1 and 2”), but they do need to say something about a 1 unit or degree change in temperature. Address the student directly."
},
{
"role": "user",
"content": "Here is the student's response: [RESPONSE HERE]"
}]

Sample Incorrect Participant Responses and Corresponding Feedback

Table S15

Practice Question 2: What does it mean that the intercept is 5.58?

Incorrect Response	AI Feedback
It intersects the Y axis at 5.58	Incorrect. The intercept does occur at 5.58 on the y-axis, but this explanation doesn't tell us what that means in the context of the data. Corrected version: It means that when the average winter temperature is zero degrees, the predicted murder rate is 5.58.
Change in y over change in the value of the x and y coordinate	Incorrect. The intercept is not a description of the change in values — that describes the slope. The intercept represents the predicted value of the y-variable (murder rate) when the x-variable (temperature) is zero. Corrected version: The intercept tells us the predicted murder rate when the temperature is zero.
Thats where the line crosses the y intercept and it means the murder rate it 5.58 when the temps are at 1.18	Incorrect. The intercept tells us the predicted value of the murder rate when the temperature is **zero**, not when it's 1.18. 1.18 is the slope, not the temperature value at the intercept. Corrected version: "That's where the line crosses the y-axis, and it means the murder rate is 5.58 when the temperature is zero."

Table S16

Practice Question 3: What does it mean that the slope for temperature is 1.18?

Incorrect Response	AI Feedback
It means that the murder rate is 5.58 when the temp is 1.18.	Incorrect. The slope tells us how much the murder rate changes when the temperature increases by 1 degree, not the value of the murder rate at a particular temperature. Corrected version: It means that the murder rate increases by about 1.18 for every 1 degree increase in temperature.
The rate at which murders increase	Incorrect. The answer is too vague and does not mention how the murder rate changes in response to a one-unit change in temperature. Corrected version: the rate at which the murder rate increases when the temperature increases by 1 degree.
The murder rate is 1.18 when the temperature is equal to zero.	Incorrect. The student's answer confuses the slope with the intercept. The slope describes how the murder rate changes with temperature, not the rate when the temperature is zero. Corrected version: The murder rate increases by about 1.18 for every 1-degree increase in temperature.

Evaluation of AI Feedback

Accuracy of AI Feedback

To validate the AI feedback participants received, we evaluated all 309 AI-generated responses on two dimensions: classification accuracy (correct versus incorrect) and, for incorrect responses, feedback accuracy (whether the AI provided an appropriate correction).

Classification Accuracy. A member of our research team found that the AI correctly labeled 45 of 50 correct responses as "correct" (90% accuracy). The 5 mislabeled responses were marked "incorrect" due to poor writing that technically met our rubric's correctness criteria. In these cases, the AI provided accurate and likely helpful feedback, though it was overly harsh. More critically, the AI demonstrated even higher accuracy in identifying incorrect answers, with a 98% correct rejection rate (255 of 259 incorrect responses correctly identified). All four false positives involved partially correct but incomplete responses. For example, one participant interpreted the slope of a line as "Rise in temp is linked to a 1.18 increase in murder rate per million resident," but failed to specify that this 1.18 increase occurs for each one-unit rise in temperature—a key requirement of the scoring criteria.

Feedback Accuracy. When the AI provided feedback on incorrect responses, it offered an appropriate correction 98% of the time (249 of 255 cases). All 6 errors involved incomplete corrections on the slope interpretation problem. Specifically, the model was prompted to "edit [each participant's] exact wording with the smallest possible changes to make their response fully correct," but it occasionally omitted part of the correct solution in doing so. For example, when a participant wrote that a slope "means that at 1.18 temperature, the murder rate will increase," the model provided this correction: "It means that for each 1 degree increase in temperature, the murder rate increases." This correction omitted the magnitude of the predicted increase (1.18 murders per million residents), a requirement in our scoring criteria.

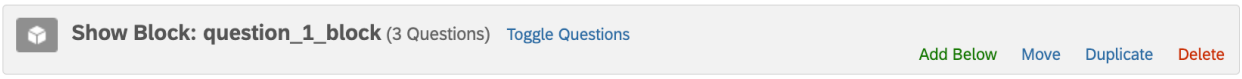
Overall, with a 90% hit rate, 98% correct rejection rate, and 98% feedback accuracy—combined with zero hallucinations (all feedback was grounded in participants' actual responses and our scoring criteria)—AI provided reliable formative feedback that was sufficiently accurate for our research purposes.

Guide for Setting Up Practice with AI Feedback

Below is a summary of how we set up AI feedback in Study 2, for researchers or educators who are interested in adopting this approach. A Qualtrics survey template that you can copy and use is available in the OSF repository for this project: https://osf.io/kruzw/files

1. Participants answered open-ended practice questions.

In our Qualtrics survey, participants typed their responses into text boxes. Each question was presented in a separate Qualtrics "block" than the question.



2. Their responses were sent to GPT-4o for evaluation.

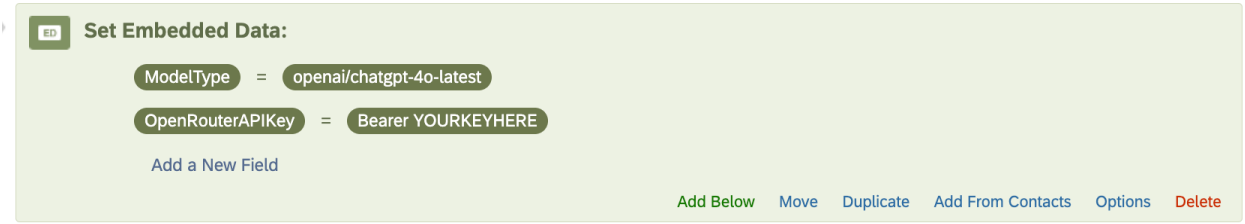
We used a service called OpenRouter (https://openrouter.ai) to communicate with GPT-4o. OpenRouter is an intermediary service that provides unified access to multiple AI models (including models from OpenAI, Anthropic, Google, and others), which made it easy to test our feedback prompts across different models during development.

To get started with OpenRouter:

- Create an account at https://openrouter.ai
- Add credits to your account (we recommend adding $5-10 at a time to minimize losses if your API key is accidentally exposed)
- Generate an API key (it will be a long string starting with `sk-or-v1-`)
- Store this key securely—you'll use it to authenticate API requests

Instead of OpenRouter, you can also obtain API keys directly from AI providers, which offer similar pricing and allow you to set spending limits.

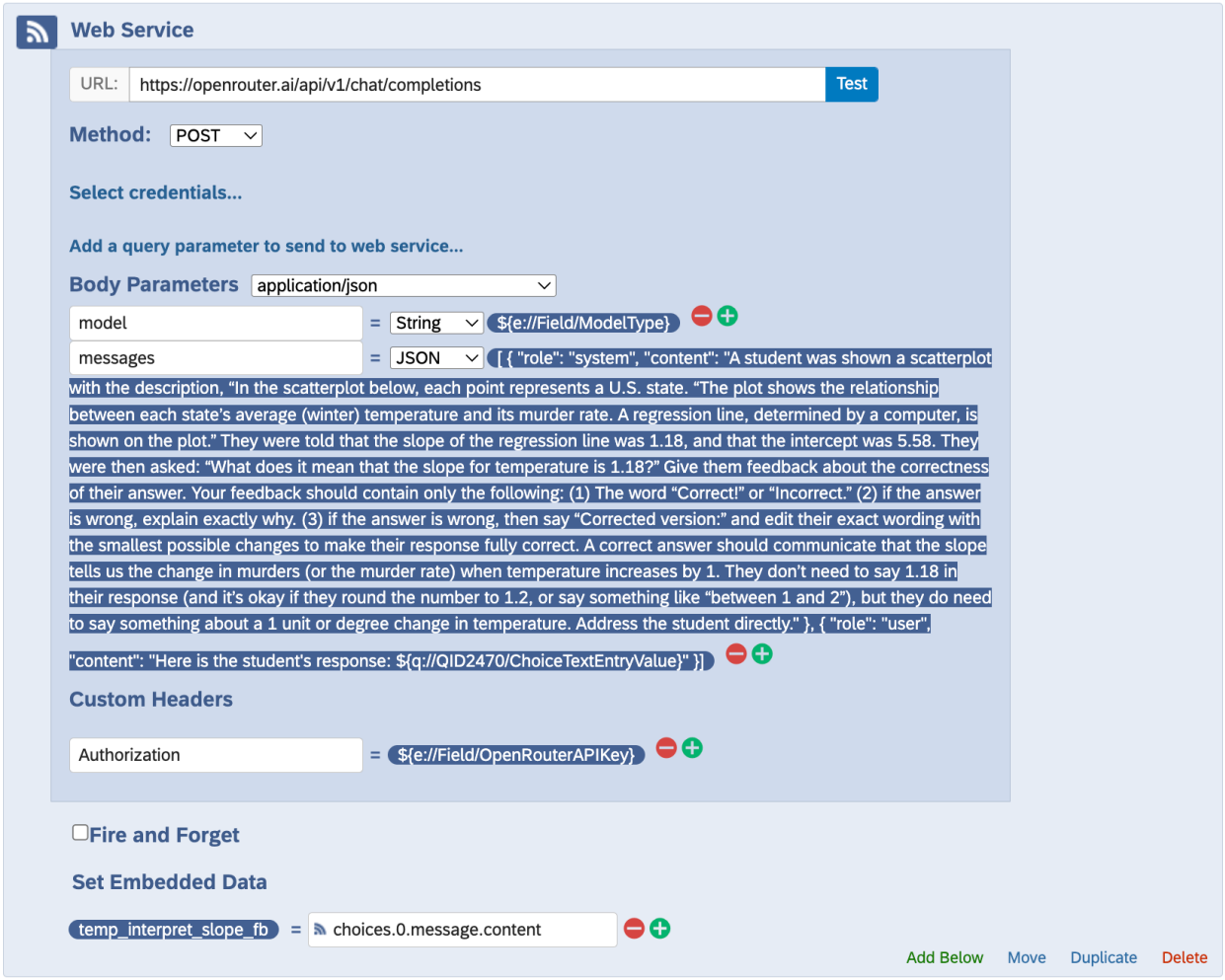
On Qualtrics, we store the API key and model name as embedded data at the beginning of the survey:



If you use this procedure, replace `YOURKEYHERE` with your key.

3. The AI evaluates the response using a pre-written prompt

Each practice question had an associated prompt that we sent to the AI along with the student's response. We interacted with the OpenRouter API using a Qualtrics “web service” block. This is what it looks like in the survey flow.



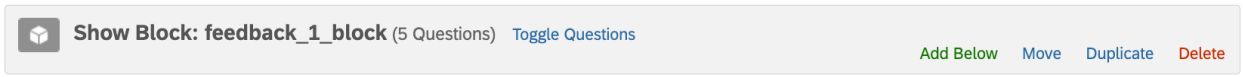
All prompts followed a consistent structure:

- A student was shown [PROBLEM CONTEXT].
- The student was then asked [QUESTION].
- Give them feedback about the correctness of their answer.
- Your feedback should contain: (1) The word "Correct!" or "Incorrect."
- (2) if the answer is wrong, explain exactly why. (3) if the answer is wrong, say "Corrected version:" and edit their exact wording with the smallest possible changes to make their response fully correct.
- A correct answer should [CORRECTNESS CRITERIA].
- Address the student directly.
- Here is the student's response: [RESPONSE]

We used piped text to pass the students' response, model type, and API key to OpenRouter. The "Set Embedded Data" section at the bottom of the block allows Qualtrics to store the LLM's response to the participant in an embedded data variable. In the example above we call this variable `temp_interpret_slope_fb`.

3. We displayed the LLM's feedback to the participants.

In the survey flow, immediately after the Web Service block with the API call, we inserted a block to display the feedback:

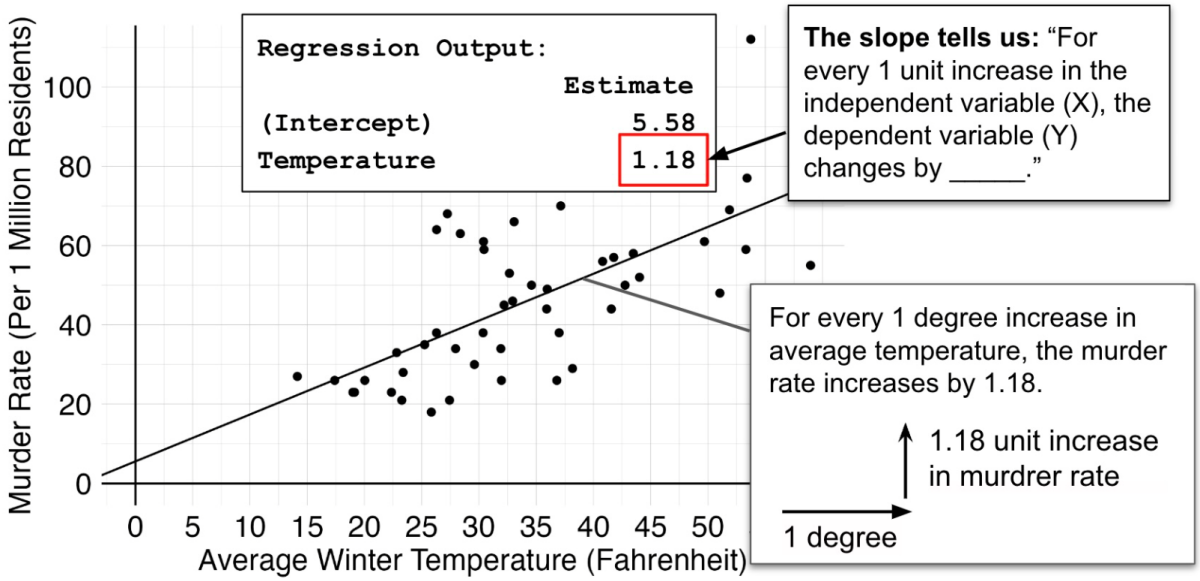


In this block we used piped text to display the students' response to the question and the AI feedback. We also included a generic explanation.

Q2577

When asked "What does it mean that the slope for temperature is 1.18?" your response was:
${q://QID2470/ChoiceTextEntryValue}

AI Feedback:
${e://Field/temp_interpret_slope_fb}



References

Miles, A. (2016). Obtaining predictions from models fit to multiply imputed data. *Sociological Methods & Research, 45*(1), 175–185. https://doi.org/10.1177/0049124115610345

Rubin, D. B. (1987). *Multiple Imputation for Nonresponse in Surveys* (1st ed.). Wiley. https://doi.org/10.1002/9780470316696

van Buuren, S., & Groothuis-Oudshoorn, K. (2011). mice: Multivariate imputation by chained equations in R. *Journal of Statistical Software, 45*(3). https://doi.org/10.18637/jss.v045.i03