Date 11-09-2026

PSYCH 10 · Fall 2026 · Practice exam · Answer key

The answers are in blue. Where a question has more than one defensible answer, the key says so and describes what a good answer contains. Each key says why the answer is right, why the tempting wrong answers are wrong, and ends with the key idea to carry to new situations.

Question 1 · Gym visits and heart rate

A gym recorded, for some of its members, the number of times each visited in a month (from 2 to 14 visits) and each member's resting heart rate, in beats per minute. Each member chooses their own number of visits. Below is the R output for a regression model.

```
hr_model <- lm(HeartRate ~ Visits, data = gym)
hr_model
Coefficients:
(Intercept)       Visits
       80.0         -2.5
supernova(hr_model)
Analysis of Variance Table (Type III SS)
Model: HeartRate ~ Visits
             SS  df       MS      F    PRE      p
Model    992.250   1  992.250 27.000 0.3600  .0000
Error   1764.000  48   36.750
Total   2756.250  49   56.250
```

1.1 A bigger branch

The gym's other branch fits the same model to four times as many of its members and also finds PRE = .36. Would that branch's F be larger than 27, smaller than 27, or about the same? Explain your answer in terms of the numerator and the denominator of F.

Answer key

Answer: *Larger (about 111). F is the ratio of the share of error the model explains, per model degree of freedom, to the share of error left unexplained, per error degree of freedom. With the same PRE, the model still uses one degree of freedom, but the error degrees of freedom grow from 48 to about 198, so the unexplained share is divided by about four times as many degrees of freedom, and F is about four times as large. The same PRE resting on four times as many members is stronger evidence that visits and heart rate are related.*

Common mistakes: *"About the same, because the PRE is the same", which ignores the amount of data behind the PRE. "Smaller, because more members means more error": SS Error grows, but so does df Error.*

Rubric (4 points): *"larger" (1). Reason (3 points): with the same PRE, more members mean more error degrees of freedom, so F rises (2); the same share of error explained is stronger evidence when it rests on more data (1). "About the same, because the PRE is the same": 0.*

Key idea: *PRE is the share of the error that a model removes; F also takes into account the amount of data behind that share. The same PRE is more convincing when it comes from more data.*

1.2 Reading the table

(i) How many members' data were used to fit this model? Name the table entry you used, and show how you got from that entry to the number of members.

(ii) The table shows df Model = 1. Which of the model's two coefficients does that one degree of freedom belong to? Why not the other coefficient?

Answer key

Answer: *(i) 50, from df Total. df Total is the number of data points minus 1, because the empty model uses one degree of freedom to estimate the mean, so n = 49 + 1 = 50. (ii) The slope, b_1 = —2.5, the coefficient on Visits. Compared with the empty model, which estimates only the mean, the regression model estimates one more parameter, the slope on visits, and that parameter uses one degree of freedom; df Error, 48, is the remainder.*

Also right: *For (ii), "the explanatory variable, visits", if the answer says that the model estimates one number for that variable.*

Common mistakes: *For (i), answering 49 (df Total) or 48 (df Error). For (ii), "the intercept": the empty model already estimates one number, and df Total has already taken it into account.*

Rubric (3 points): *(i) 1.5 points: 50 (0.5), from df Total plus 1 (1). (ii) 1.5 points: the slope, one parameter beyond the empty model (1.5). "The intercept": 0.5 of 1.5.*

Key idea: *The df column counts the parameters a model has used: one for the empty model's mean, and one more for each added parameter. Read in reverse, the column gives the number of data points behind the table.*

1.3 The AI assistant's summary

The gym asked an AI assistant to describe the output. Below is its summary, one claim per sentence. AI assistants can write summaries that sound convincing but contain mistakes.

[1] Members who visit more tend to have lower resting heart rates: each additional visit per month is associated with a decrease of about 2.5 beats per minute in resting heart rate. **[2]** So a member who adds four visits a month will lower their resting heart rate by 10 beats per minute. **[3]** A member who never visits would have a resting heart rate of 80. **[4]** Visits explain 36% of the variation in resting heart rate across members. **[5]** Because members chose how often to visit, these results tell us nothing about how visits and heart rate are related.

Rate each claim by how well the R output above justifies the claim: well, partly, or not at all. For each claim rated "partly" or "not at all", explain the reasoning behind your rating.

Answer key

Well justified: *[1], the slope $b_1 = -2.5$, read as an association. [4], PRE = .36.*

Not justified:

- ***[2]*** *Members chose their own number of visits, so these are observational data. Members who visit more may differ in other ways, such as age, diet or fitness before joining, so the slope describes a difference between members, not the effect of a change in one member's visits.*
- ***[3]*** *The intercept is the line's value at 0 visits, but every member visited at least twice. Predicting at 0 visits extends the line beyond the range of the data, where the data give no support for the line.*
- ***[5]*** *Self-selection limits the causal conclusions the data support, not their description of these members. The slope and PRE still describe the association between visits and heart rate among these members, as in claims [1] and [4].*

Also right: *"Partly" for [2] or [3], with the reasons above (the slope does describe an association; the intercept is the line's value at 0). "Partly" for [5], with the reason that the data say nothing about cause but do describe the association.*

What makes it strong: *A strong answer checks each claim against the meaning of the number behind it: the slope as an association in observational data, the intercept against the range of the data, and self-selection as a limit on causal claims, not on description.*

Weaker answers: *Rating [1] "not at all" because "correlation isn't causation" overlooks that [1] claims only an association. Rating [5] "well" makes the opposite mistake: it treats a limit on causal claims as a limit on description. Rating [4] "not at all" misreads PRE.*

Rubric (5 points): *1 point per claim: [1] and [4] well (0.5 for the rating, 0.5 for a correct reading); [2], [3] and [5] "not at all" or "partly" (0.25) with the reason given above (0.75). A "partly" with a reason that supports the claim: 0.25 of 1.*

Key idea: *Check each claim in an AI summary separately, against the meaning of the number behind it. Summaries most often err on claims about causes and on values outside the range of the data, and the errors run in both directions, overstatement and understatement.*

Question 2 · Background noise and reaction time

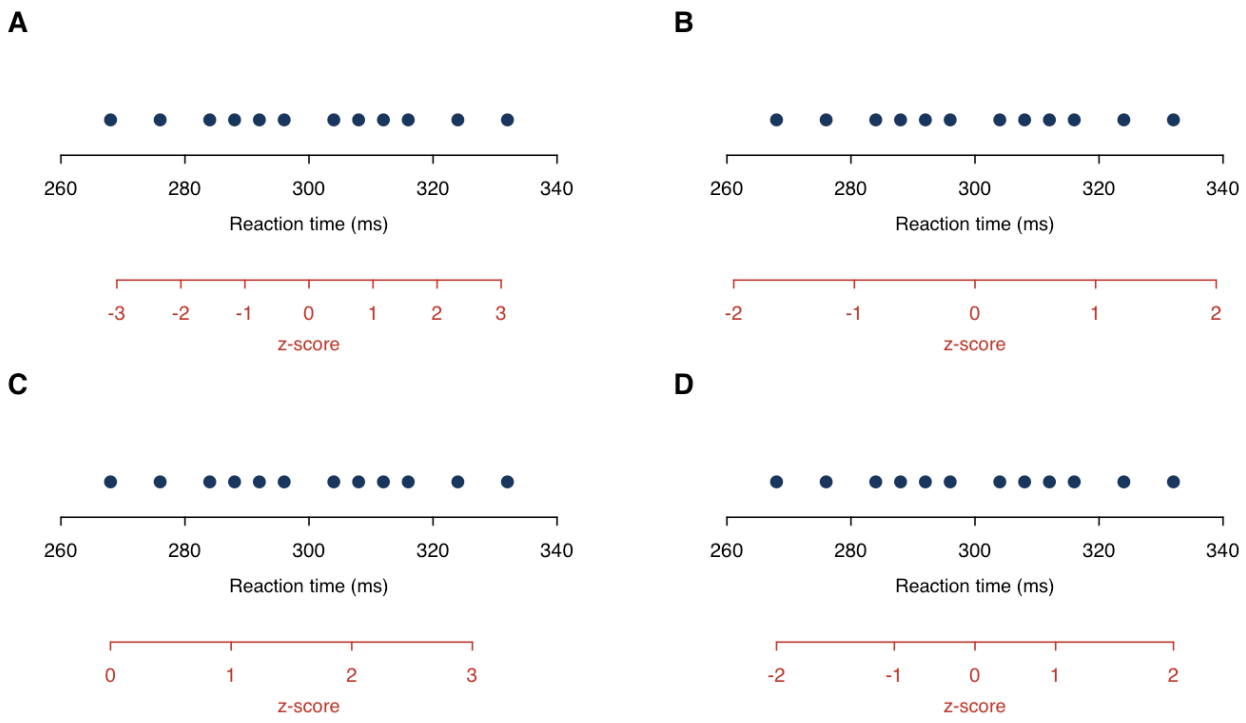
A class randomly assigned 12 volunteers to do a reaction-time task in one of three rooms, 4 volunteers per room: quiet, low background noise, or loud background noise. Below is the R output for reaction time, in milliseconds (ms), overall and by room.

```
favstats(~ RT, data = noise)
 min  Q1 median  Q3 max mean       sd  n missing
 268 287    300 313 332  300 19.44694 12       0
favstats(RT ~ Room, data = noise)
   Room min Q1 median  Q3 max mean       sd n missing
1  Loud 308 314    320 326 332  320 10.32796 4       0
2   Low 288 294    300 306 312  300 10.32796 4       0
3 Quiet 268 274    280 286 292  280 10.32796 4       0
```

The empty model predicts 300 ms, the overall mean, for every volunteer; its SS, the sum of squared errors, is about 4,160.

2.1 The slowest volunteer

(i) Each panel shows the 12 reaction times with a second axis labeled "z-score". Which panel's second axis correctly shows each reaction time's z-score, computed from the mean and standard deviation of all 12 volunteers? Choose one, and give your reason.



(ii) The slowest volunteer, at 332 ms, was in the Loud room. About how many standard deviations above the mean is 332 ms, using the mean and standard deviation of all 12 volunteers? And using the mean and standard deviation of the 4 volunteers in the Loud room? A rough value is enough for each (for example, "about 0.5"). Which of the two values is larger?

(iii) Which of the two comparisons in (ii) is the right one for judging how unusual this time is for a volunteer in the Loud room? Explain the reason for the difference between the two values in (ii).

Answer key

Answer: *(i) B. A z-score is a value's distance from the mean, in standard deviations, with a sign. On axis B, 0 sits at the mean, 300 ms, and each unit is one standard deviation, about 19 ms, so 268 ms is at about —1.6 and 332 ms at about +1.6. (ii) Among all 12: 332 is 32 ms above 300, about 1.6 standard deviations of about 19 ms. Among the Loud room: 12 ms above 320, about 1.2 standard deviations of about 10 ms. So the larger value is the one for all 12. (iii) The Loud-room comparison is the right one: it judges the time against the other volunteers in the same room. The two values differ because each compares the time with a different distribution: the Loud room's mean is higher, so 332 is closer to that mean, and the room's standard deviation is smaller, since much of the overall spread comes from the differences between rooms.*

Why not the others in (i):

- *A centers correctly but uses a room's standard deviation, about 10 ms, instead of all 12 volunteers' standard deviation, so the units are too small.*
- *C starts at 0 at the fastest time, as if a z-score were a distance that cannot be negative. Times below the mean have negative z-scores.*
- *D labels the quartiles —2 to 2, as if z-scores were ranks. z-scores are measured in standard deviations, so they are evenly spaced in ms, unlike these labels.*

Common mistakes: *For (ii), dividing by the variance instead of the standard deviation, or using the overall mean with the Loud room's standard deviation.*

Rubric (5 points): *(i) 2 points: B (0.5), because 0 is at the mean and one unit is the overall standard deviation (1.5). (ii) 1 point: about 1.6 and about 1.2, the larger value for all 12 (1). (iii) 2 points: the Loud-room comparison (0.5); the two values differ because the Loud room has a higher mean and a smaller standard deviation (1.5).*

Key idea: *A z-score always answers "compared with what?". The same value can be unusual in one distribution and ordinary in another, so choose the distribution that matches the question.*

2.2 One volunteer drops out

The class learns that one volunteer in the Low room misread the instructions. That volunteer's time was 296 ms, and the class removes it from the data. Compared with the standard deviation of all 12 times, about 19.4 ms, is the standard deviation of the 11 remaining times larger, smaller, or about the same? Choose one, and give your reason.

A. Smaller, because the total of squared errors now has one fewer term, so the total is smaller.
B. Larger, because 296 is near the mean: the total of squared errors barely changes, but it is now divided by 10 instead of 11.
C. About the same, because removing one volunteer out of 12 makes little difference to the spread of the times.
D. Smaller, because a smaller group of volunteers has less variation among its times, so the summary of that variation is smaller.

Answer key

Answer: *B.*

Why: *The standard deviation comes from the total of squared errors divided by n — 1. The time 296 ms is only 4 ms from the mean, so its squared error, 16, is small: removing it takes the total from about 4,160 to about 4,140. But the divisor drops from 11 to 10, so the variance rises from about 378 to about 414 and the standard deviation from about 19.4 to about 20.4 ms. Removing a time far from the mean, such as 332, would make the standard deviation smaller.*

Why not the others:

- *A treats the standard deviation as a total. It is a per-case measure, so it depends on the divisor as well as on the total.*
- *C ignores that one value can move the standard deviation noticeably in a dataset this small.*
- *D assumes that fewer cases always mean less spread. Spread depends on the distances of the remaining values from their mean, not on the number of values.*

Rubric (4 points): *B (1 point). Reason (3 points): 296 is near the mean, so the total of squared errors barely changes (1.5); the divisor n — 1 drops, so the standard deviation rises (1.5). A with "one fewer error in the total": 0.5 of 3.*

Key idea: *SS is a total; the variance and standard deviation are per-case measures, so they depend on whether a removed value was typical or unusual.*

2.3 What the rooms explain

The Room model, `lm(RT ~ Room)`, predicts each volunteer's reaction time with the mean for their room instead of 300 ms.

(i) What would the Room model's PRE be if the Loud, Low and Quiet rooms each had a mean of exactly 300 ms? What would it be if every volunteer's time equaled the mean of their room? Give each number and explain your reasoning.

(ii) An AI assistant summarized the Room model in the three sentences below. Rate each sentence by how well the R output above justifies the sentence: well, partly, or not at all, and explain the reasoning behind each rating.

[1] Of all models that make one prediction per room, the model that predicts each room's mean has the smallest total squared error. **[2]** The Room model's predictions are exactly right for every volunteer in the Low room. **[3]** Because the Room model's SS Error is smaller than the empty model's, the three rooms must differ in their effect on reaction time.

Answer key

Answer: *(i) 0 and 1. If every room's mean were 300, the Room model would make the same predictions as the empty model, remove none of its error, and PRE would be 0. If every time equaled its room's mean, the Room model would leave no error at all, so it would remove all of the empty model's error, and PRE would be 1. (ii) [1] well; [2] and [3] not at all.*

- ***[1]*** *Within each room, the room's mean is the single prediction that leaves the smallest total of squared errors, so no other set of three predictions could leave a smaller total.*
- ***[2]*** *The model predicts 300 ms for every Low-room volunteer, but their times run from 288 to 312 ms. The room's standard deviation, about 10 ms, is the typical size of the prediction errors within the room.*
- ***[3]*** *Adding groups to a model always reduces SS Error or leaves it unchanged, even when the groups have no real difference, because each group's mean follows the chance variation in its own data. F and its p-value compare the reduction with the size of reduction expected from chance alone.*

Also right: *"Partly" for [3], with the reason that the rooms may well differ but a smaller SS Error alone does not show a difference.*

Common mistakes: *For (i), "100" for the second case (PRE is a proportion, at most 1). Rating [1] "not at all" because "the median might do better": the mean is the prediction that minimizes squared error.*

Rubric (5 points): *(i) 2 points: 0 (0.5) because the model removes none of the error (0.5); 1 (0.5) because it removes all of it (0.5). (ii) 3 points: [1] well, with the reason that the mean minimizes squared error (1); [2] not at all, because times vary within the room (1); [3] not*

at all or partly, because adding groups always reduces SS Error, so a reduction alone is not evidence (1).

Key idea: *PRE runs from 0, for a model no better than the empty model, to 1, for a model with no error left. A group model predicts each group's mean, the prediction with the smallest squared error within the group, so a group model always fits at least as well as the empty model. The question for inference is whether its improvement exceeds the improvement expected from chance alone.*

2.4 Checking a classmate's work

(i) A classmate finds the empty model's SS by squaring the overall standard deviation and multiplying by 12, the number of volunteers. Is the classmate's SS too large, too small, or correct? Explain your answer; you do not need to compute the SS.

(ii) Another classmate says, "The standard deviation of about 19 ms means that each volunteer's time is about 19 ms from the mean of 300." How well justified is that statement: well, partly, or not at all? Explain the reasoning behind your rating.

Answer key

Answer: *(i) Too large. The variance is SS divided by n — 1, so SS is the variance times n — 1 = 11, not times 12. Multiplying by 12 gives a total one-eleventh too large (about 4,540 instead of about 4,160). (ii) Partly. About 19 ms is a typical size of the empty model's errors, but individual volunteers' distances from 300 vary widely: the 296-ms and 304-ms volunteers are 4 ms away, and the 268-ms and 332-ms volunteers are 32 ms away. The standard deviation summarizes the errors as a group; it does not give each volunteer's distance from the mean.*

Also right: *For (i), computing both totals ($378 \times 12 \approx 4{,}540$ against $378 \times 11 \approx 4{,}160$) is fine but not needed. For (ii), "not at all", with the reason that no volunteer need be exactly 19 ms from the mean.*

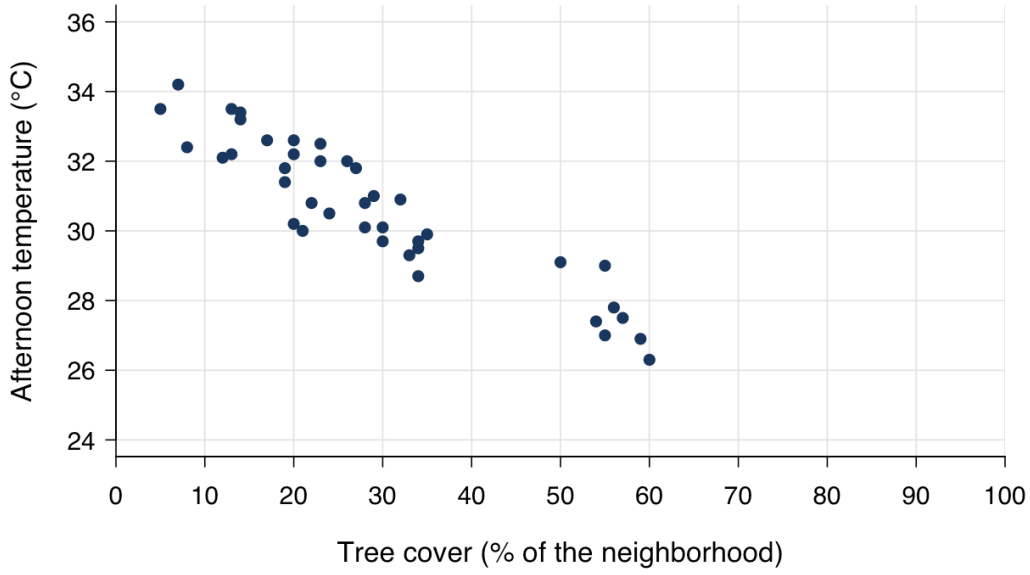
Common mistakes: *For (i), "correct, because there are 12 volunteers". For (ii), "well, because that is what a standard deviation is".*

Rubric (4 points): *(i) 2 points: "too large" (0.5), because SS is the variance times n — 1, not n (1.5). (ii) 2 points: "partly" or "not at all" (0.5); the standard deviation is a typical size of the errors, and individual distances vary (1.5). "Well": 0.*

Key idea: *The standard deviation in the output is a summary of the empty model's error: squaring it and multiplying by n — 1 recovers the total, and it describes a typical error, not every error.*

Question 3 · Trees and summer heat

A city measured tree cover (the percent of a neighborhood shaded by trees) and the mean summer afternoon temperature in 40 of its neighborhoods. Here are the data.



3.1 Which prediction?

A planner fits a straight line to the 40 points (the line is not drawn on the plot). At 21% tree cover the line gives about 31.5 °C, and the one neighborhood that was measured at 21% cover had a mean temperature of 30.0 °C. The planner considers four predictions. Which one is best supported by these data? Choose one, and give your reason.

A. A new neighborhood in this city with 21% cover will be about 30.0 °C, the same as the one measured neighborhood with 21% cover.
B. A new neighborhood in this city with 21% cover will be about 31.5 °C, the line's value at 21%.
C. A planned neighborhood in this city with 95% cover will be about 23 °C, the line's value extended to 95%.
D. A neighborhood with 21% cover in a distant desert city will be about 31.5 °C, the line's value at 21%.

Answer key

Answer: *B.*

Why: *A line that follows the pattern of all 40 neighborhoods draws on every measurement, so its value at 21%, about 31.5 °C, is a better prediction for an unmeasured neighborhood than the temperature of any single neighborhood. Neighborhoods scatter about 1 °C above and below the line, and the one measured at 21% happens to lie about 1.6 °C below the line.*

Why not the others:

- *A rests on one neighborhood, which carries its own chance variation; a second neighborhood at 21% would probably have a different temperature.*
- *C extends the line far beyond the data: the most shaded neighborhood measured has about 60% cover, and nothing in the data says the pattern continues to 95%.*
- *D applies the pattern to a place the 40 neighborhoods do not represent. A desert city's climate, buildings and trees may relate in a different way.*

Rubric (4 points): *B (1 point). Reason (3 points): the line uses all 40 neighborhoods, so it is a better guide than one neighborhood (1); C is beyond the range of the data (1); D concerns a city not represented in the data (1). An answer that chooses B and explains only why A is worse: 2 of 3 for the reason.*

Key idea: *Predict from the overall pattern, not from the one case that happens to match, and only for values and places like the ones the data come from.*

3.2 The hills

In this city, the neighborhoods with the most trees are in the hills, which are higher than the rest of the city. A council member looks at the plot and says, "Planting trees cools neighborhoods."

(i) Suppose trees had no effect on temperature. How would elevation have to be related to temperature for the plot to show this same downward pattern? Explain your answer.

(ii) How well do these data support the council member's claim: well, partly, or not at all? Explain the reasoning behind your rating, using your answer to (i).

Answer key

A strong answer: *(i) Higher neighborhoods would have to be cooler. The description of the city already ties elevation to tree cover (the leafiest neighborhoods are in the hills); if elevation is also tied to temperature, the leafy neighborhoods would be the cool ones even if the trees had no effect, and the plot would look the same. (ii) Partly, or not at all. The data are consistent with trees cooling neighborhoods, but if higher places are cooler, some or all of the pattern could come from elevation, so these data alone cannot show that planting trees would cool a neighborhood. An experiment, such as one that assigns trees to places at random, would settle the question.*

What makes it strong: *A strong answer sees that elevation must be tied to both tree cover and temperature to produce the pattern, and concludes that the data alone cannot establish cause.*

Also right: *For (i), "the hills would have to be cooler for some other reason than their trees, such as their height". For (ii), either "partly" or "not at all", with the elevation reason.*

Weaker answers: *For (i), "elevation affects trees" without the link to temperature. For (ii), "correlation isn't causation" without an account of how elevation could produce the pattern, or "well" because the downward trend is clear.*

Rubric (4 points): *(i) 2 points: higher neighborhoods would have to be cooler (1.5), so elevation alone would give the same pattern (0.5). (ii) 2 points: "partly" or "not at all" (0.5), because the data cannot separate trees from elevation (1.5). "Well": 0.*

Key idea: *A variable tied to both tree cover and temperature can create a relationship between them even if neither affects the other. Before reading a pattern as cause, ask which other variables differ between the cases.*

3.3 The cool-pavement lottery

Trees take years to grow, so the city wants to test a quicker alternative: a reflective coating for streets. It can afford to coat 30 streets this summer, and residents of 60 streets have asked for the coating. A planner proposes a lottery: 30 of the 60 are drawn at random to be coated this summer, and the other 30 are coated next year. Over the summer, the city will compare afternoon temperatures on the two sets of streets.

(i) Could this comparison show whether the coating cools streets? Explain your reasoning.

(ii) A resident objects: "A lottery is unfair; you're experimenting on people." How well justified is this objection: well, partly, or not at all? Judge whether the lottery is fair to these residents, and explain the reasoning behind your rating.

Answer key

A strong answer: *(i) Yes. The lottery decides which streets are coated, so elevation, shade, traffic and every other factor that affects temperature, including factors nobody measured, are spread about evenly across the two sets of streets by chance. A clear difference in temperature can then be attributed to the coating. (ii) Not at all. The city can coat only 30 streets this summer either way, so the lottery takes nothing away from anyone; it decides fairly the order in which the streets are coated, and every street that asked is coated within a year. The coating is not expected to harm anyone, and the residents requested the coating. A design that withheld something people need, or exposed them to harm, would be a different matter even if it were statistically ideal.*

What makes it strong: *(i) explains that random assignment balances other factors, known and unknown. (ii) gives the reasons the lottery is acceptable here: scarcity, every street coated eventually, no harm.*

Also right: *"Partly", with the reason that the residents should be told about the study and its purpose, as long as the answer explains why the lottery itself is fair.*

Weaker answers: *"Because they will measure more carefully" overlooks the role of the lottery. "Any experiment on people is unethical" or "it is fine because it is random" give no reason specific to this design.*

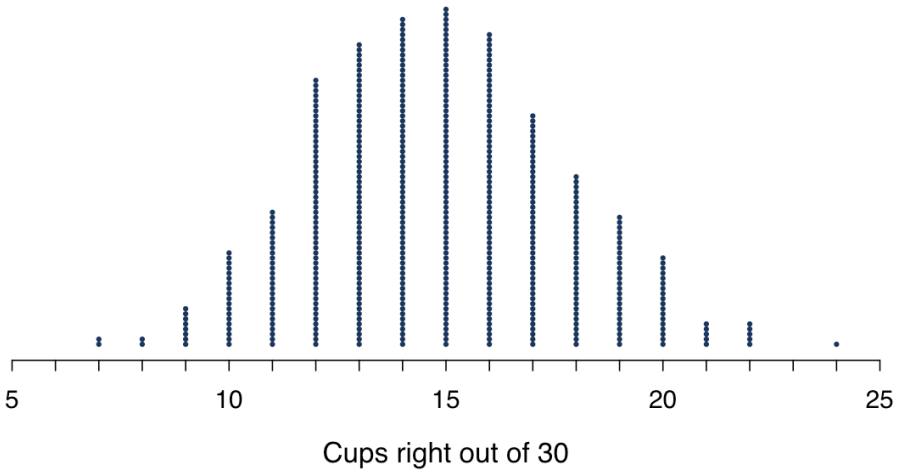
Rubric (5 points): *(i) 2.5 points: random assignment spreads other factors, known and unknown, evenly across the two sets of streets (2), so a difference can be attributed to the coating (0.5). (ii) 2.5 points: "not at all" or "partly" (0.5), with the reasons: scarcity, every street coated within a year, no harm (2, any two). "Well justified": 0.*

Key idea: *Random assignment is the basis for a causal conclusion. When a beneficial treatment is scarce, a lottery is often the fairest way to allocate the treatment and, at the same time, an opportunity for an experiment.*

Question 4 · Telling fresh coffee from stale

A barista says she can taste whether a cup of coffee was made from beans roasted this week or last month. To test this claim, a friend prepares 30 cups. For each cup the friend flips a coin to decide which beans to use, and the barista tastes the coffee, without seeing the beans, and says whether the beans were fresh or old. She is right for 19 of the 30 cups.

To judge this result, the friend ran a computer simulation of a taster who only guesses: for each of the 30 cups, the taster is right with probability one half. The friend repeated this simulation 500 times and made the plot below. Each dot is one simulated guesser, placed at the number of cups out of 30 that guesser got right.



4.1 Nineteen of thirty

(i) From the plot, estimate the number of the 500 simulated guessers who got 19 or more cups right (to the nearest 10 is fine), and say about how many that is out of every 100 guessers.

(ii) The friend says, "So this shows she can't really tell the difference." How well justified is the friend's conclusion: well, partly, or not at all? Explain the reasoning behind your rating.

Answer key

A strong answer: *(i) About 55 of the 500 (the columns at 19, 20, 21, 22 and 24 hold 26, 18, 5, 5 and 1 dots), roughly 11 in every 100; any count from about 40 to 70, or 8 to 14 in 100, is right. (ii) Not at all. A guesser scores 19 or more about 11 times in 100, which is not rare, so 19 of 30 is weak evidence that she can taste the difference. But the result does not show that she* cannot*: a taster who tells the difference somewhat better than chance could easily score 19 of 30 as well. The result cannot distinguish guessing from a modest skill; only a longer test could make that distinction.*

What makes it strong: *A strong answer reads the share of simulated results at least as extreme as hers from the plot, and separates "not convincing" from "shown to be guessing".*

Also right: *"Partly", with the reason that the result is indeed weak evidence of skill, as long as the answer says weak evidence is not evidence of no skill.*

Weaker answers: *"19 is more than 15, so she can tell" ignores the frequency with which a guesser reaches 19. "Well justified, because 11 in 100 isn't rare" treats a lack of evidence for skill as evidence against skill.*

Rubric (5 points): *(i) 1.5 points: a count of about 40 to 70 (1), as about 8 to 14 in every 100 (0.5). (ii) 3.5 points: "not at all" or "partly" (0.5); guessing reaches 19 fairly often, so the result is weak evidence (1.5); weak evidence of skill is not evidence of no skill (1.5). "Well justified": 0.5 of 3.5 for using the rate.*

Key idea: *Judge a result by the frequency with which chance alone produces a result at least as extreme. A result that chance produces often is weak evidence, which is not the same as evidence of no skill.*

4.2 A longer test

The barista takes a longer test: 120 cups, chosen by coin flips in the same way. She gets 76 right, again about 63%. How does this result compare with her 19 of 30, as evidence that she can taste the difference? Choose one, and give your reason.

A. About the same evidence, because she got the same share of cups right both times.
B. Weaker evidence, because with 120 cups she had more chances to make mistakes.
C. Stronger evidence, because a guesser's share stays nearer one half on a longer test.
D. Stronger evidence, because she got 57 more cups right than she did on the first test.

Answer key

Answer: *C.*

Why: *On 30 cups, a guesser's share of correct answers varies widely by chance: about 11 in 100 guessers reach 63%. On 120 cups, lucky and unlucky cups offset one another far more, so a guesser's share stays much closer to 50%, and a share of 63% by guessing becomes rare (about 2 in 1,000). The same 63% now lies far out in the tail of the guessing distribution.*

Why not the others:

- *A judges the evidence by the share alone. The rarity of a share under guessing depends on the number of cups behind that share.*
- *B counts mistakes rather than the share right. She made more mistakes, but out of many more cups; a guesser's mistakes would also grow with the number of cups.*
- *D compares counts from tests of different lengths. A guesser would also get many more cups right on 120 cups than on 30; the relevant quantity is the distance of her share from a guesser's share.*

Rubric (3 points): *C (1 point). Reason (2 points): a guesser's share varies less on more cups, so 63% is much rarer by chance on 120 (2). D, or "more data is always better" alone: 0.5 of 2.*

Key idea: *The same percentage is stronger evidence when it rests on more observations, because chance variation in a percentage is smaller for a larger sample.*

4.3 The tasting contest

A chain of coffee shops holds a contest. Each of its 40 baristas tastes 30 cups prepared in the same way. The chain will call a barista "gifted" if the barista gets 20 or more cups right. A guesser gets 20 or more right about 6 times in 100 (29 of the 500 simulated guessers in the plot).

(i) If all 40 baristas were only guessing, about how many would you expect the chain to call gifted? Give the number and show the calculation.

(ii) Three baristas score 20 or more. Is that more than the number expected from 40 guessers? What do the three scores show, if anything, about whether any of the 40 baristas is gifted?

Answer key

A strong answer: *(i) About 2. Each guesser has about a 6-in-100 chance of scoring 20 or more, and there are 40 guessers: $40 \times 6/100 \approx 2.4$. (ii) Only slightly more than expected. Three is close to the 2 or so expected from 40 guessers, and a contest of 40 guessers often produces three or more such scores. So the three scores alone are weak evidence that any barista has a gift, and they cannot identify which of the three, if any, is gifted. Testing many baristas makes a few high scores likely by chance alone.*

What makes it strong: *A strong answer combines the single-guesser chance with the number of baristas tested, and compares the count of high scores with the number expected from chance alone.*

Also right: *For (i), 2 or 3. For (ii), "about what you would expect", or "slightly more, but not enough to show anything".*

Weaker answers: *"Three gifted baristas, since each scored well above 15" judges each score as if it were the only score tested. "None of them can have a gift, since three is about what chance gives" treats weak evidence as evidence of no gift.*

Rubric (4 points): *(i) 1.5 points: about 2 (0.5), from $40 \times 6/100$ (1). (ii) 2.5 points: about the number expected from 40 guessers (1); so the scores are weak evidence of any gift (1); many tests make a few high scores likely by chance, or the scores cannot identify a gifted barista (0.5).*

Key idea: *Many tests produce some results that look like findings even when no effect exists. Before calling a result a discovery, compare the number of such results with the number expected from chance alone.*

4.4 The retest

The chain retests the three high scorers on 30 new cups each.

(i) If the three were only guessing, about what would you expect each to score on the retest? Give a number and explain your reasoning.

(ii) The chain sets a rule: "Anyone who scores below 20 on the retest has no gift." Consider a taster who is genuinely right on 65 of every 100 cups. In a simulation of such a taster on 30 cups, the score was below 20 about half the time. How well justified is the chain's rule: well, partly, or not at all? Explain the reasoning behind your rating.

Answer key

A strong answer: *(i) About 15, most likely between 10 and 20, since a guesser is right half the time. The three were selected because they scored high, so if their first scores came from luck, their retest scores should be lower than their first scores. (ii) Not at all. A taster with a real but modest skill scores below 20 about half the time, so a score below 20 says little about whether the barista has a gift; the rule would wrongly dismiss about half of such tasters. A real skill does not produce a high score on every test; a decision requires several tests, or a test with many more cups.*

What makes it strong: *A strong answer predicts a retest score near 15 for a guesser, explains why the high scorers are expected to score lower on the retest, and recognizes that a real skill can fall short of a cutoff on a single test.*

Also right: *For (i), a range from about 11 to 19. For (ii), "partly", with the reason that a low retest is some evidence against a strong gift but cannot rule out a modest gift.*

Weaker answers: *"About 20 again" assumes the high scores reflect skill. "Well justified, because a gifted taster should pass" assumes that a real skill produces a high score on every test.*

Rubric (5 points): *(i) 2 points: about 15 (0.5), because guessers are right half the time (0.5), with the point that high scorers selected for their scores should score lower on the retest (1). (ii) 3 points: "not at all" or "partly" (0.5), because a taster with a real but modest skill scores below 20 about half the time (1.5), so one test cannot settle the question (1). "Well": 0.*

Key idea: *A result selected because it stood out is usually smaller when the measurement is repeated, and a real effect does not appear in every study. Judge by results that recur across repetitions, not by a single test in either direction.*

Question 5 · A reading app

A company makes a reading app used by 40,000 students in 120 school districts.

5.1 Two surveys

The company emailed a satisfaction survey to all 40,000 users. Of the 4,000 who replied, 85% said they liked the app. A researcher instead chose 200 users at random and kept asking until 190 of the 200 had answered the same survey; 62% of those 190 said they liked the app.

(i) As an estimate of the percentage of all 40,000 users who like the app, is the emailed survey's 85% likely too high, too low, or about right? Explain your reasoning, based on which users chose to answer the emailed survey.

(ii) Which estimate would you trust more, the 85% or the 62%? Explain your reasoning, taking into account both the way each sample was chosen and the number of people who answered.

Answer key

A strong answer: *(i) Too high. The users who answered the emailed survey selected themselves, and students who like the app are probably more likely to reply, so the replies over-represent satisfied users. (ii) The researcher's 62%. The 200 users were chosen at random, and almost all of them answered, so they should resemble all users; a sample of 190 has some chance error, perhaps several percentage points in either direction, but no bias in a known direction. The 4,000 replies do not remove the emailed survey's bias: a larger sample with the same bias gives a more precise estimate of the wrong quantity.*

What makes it strong: *A strong answer identifies self-selection of the respondents as the problem, explains why sample size cannot remove it, and weighs that bias against the random sample's chance error.*

Also right: *For (i), "about right is possible, but we cannot know", if the answer still says that the replies are probably biased and prefers the random sample.*

Weaker answers: *"The 85%, because 4,000 is much more than 190" judges a sample by its size alone. "Neither, because 200 of 40,000 is too small a share" assumes a sample must be a large fraction of the population.*

Rubric (5 points): *(i) 2 points: "too high" (0.5), because those who chose to reply probably differ from all users (1.5). (ii) 3 points: the 62% (0.5); the random choice and high response make the 190 resemble all users (1); size reduces chance error but not the bias in who replied, so the 4,000 replies do not improve the estimate (1.5). "The 85%, because it is bigger": 0 for (ii).*

Key idea: *The way a sample was chosen matters more than its size. Size reduces chance error; it has no effect on bias.*

5.2 The district rankings

The company ranks the 120 districts by their students' mean gain on a reading test over the year. The 10 districts at the top of the ranking and the 10 at the bottom all have fewer than 30 students using the app. A sales manager says, "Small districts use the app best. Let's market it to small districts."

(i) Explain how the smallest districts could end up at both the top and the bottom of the ranking even if district size had no relation to the app's effectiveness.

(ii) How well justified is the sales manager's conclusion: well, partly, or not at all? Explain the reasoning behind your rating.

Answer key

A strong answer: *(i) A mean of a few students varies much more by chance than a mean of many students. In a district of 20 students, a few unusually strong or weak readers can move the mean by a large amount; in a district of 2,000, the same readers barely change the mean. So the small districts' means are spread more widely in both directions, and the most extreme means, high and low, come mostly from small districts. (ii) Not at all. Small districts are also at the bottom of the ranking, a pattern that chance variation predicts and the manager's explanation leaves unexplained; the ranking gives no evidence that the app works better in small districts.*

What makes it strong: *A strong answer links district size to the chance variation in a mean, and uses the fact that small districts are at* both *ends of the ranking.*

Also right: *"Partly", with the reason that small districts might do better, but these rankings cannot show it because chance alone puts small districts at both ends.*

Weaker answers: *"Small districts give students more attention" explains only the top of the ranking and ignores the bottom.*

Rubric (4 points): *(i) 2.5 points: means of fewer students vary more by chance (1.5), so small districts fill both ends (1). (ii) 1.5 points: "not at all" or "partly" (0.5), with the reason that small districts are also at the bottom (1). "Well justified": 0.*

Key idea: *Means from small samples vary more than means from large samples. When the most extreme results all come from the smallest groups, suspect chance variation before looking for another cause.*

5.3 The big trial

The company then ran a trial. It randomly assigned 20,000 students to two groups: one group used the app for a year, and the other group did not. On a 100-point reading test, the app group's mean was higher by 0.4 points. The researchers then shuffled the group labels 1,000 times, each time reassigning "app" and "no app" to the students at random and recomputing the difference between the two groups' means. No shuffled difference was as large as 0.4 points. Teachers consider a gain of about 5 points meaningful. The company's ad says, "Proven to dramatically improve reading scores!"

(i) What does the shuffle result show about whether the app has any effect on reading scores?

(ii) How well justified is the ad's claim: well, partly, or not at all? Explain the reasoning behind your rating, and say whether each of the two words "proven" and "dramatically" is justified.

Answer key

A strong answer: *(i) In 1,000 shuffles, chance alone never produced a difference as large as 0.4 points between groups of this size, so the data are strong evidence that the app has some effect; because the students were randomly assigned, the effect can be attributed to the app rather than to some other difference between the groups. (ii) Partly. "Proven to improve" is justified: the evidence for some improvement is strong. "Dramatically" is not: the improvement is 0.4 points on a 100-point test, less than a tenth of the 5 points teachers consider meaningful. With 20,000 students, even a very small effect is distinguishable from chance.*

What makes it strong: *A strong answer separates the strength of the evidence for an effect (the shuffle result) from the size of the effect (0.4 points), and compares that size with the 5 points teachers consider meaningful.*

Also right: *"Not at all", if the reason is that "dramatically" makes the claim misleading as a whole, with the evidence for some effect acknowledged.*

Weaker answers: *"No shuffled difference was that large, so the effect is large" reads the evidence as a size. "The ad is true because the result is significant" does not address "dramatically".*

Rubric (5 points): *(i) 2 points: strong evidence of some effect (1), attributable to the app because of random assignment (1). (ii) 3 points: "partly" or "not at all" (0.5); "proven" is supported by the shuffle result (1); "dramatically" is not, since 0.4 is far below 5 (1.5). "Well justified": 0.5 of 3 if (i) is right.*

Key idea: *A result that chance cannot explain is strong evidence of* some *effect, not of a large effect. A very large study can distinguish even a tiny effect from chance, so always consider the estimated size of the effect.*

Question 6 · A city bike-share

A city's bike-share system recorded every trip taken in May, 2,000 trips in all. For each trip, the system recorded the start station's number (an ID from 1 to 80 printed on each station), the neighborhood where the trip started, the trip's length in minutes, and the temperature when the trip started.

6.1 The analyst's two models

An analyst fits two regression models to the 2,000 trips and writes a finding from each:

- **(a)** From `lm(Minutes ~ StationNumber)`, with slope 0.05: "Trips that start at a station numbered one higher are, on average, 0.05 minutes longer."
- **(b)** From `lm(Minutes ~ Temperature)`, with slope 0.3: "Trips that start when it is one degree warmer are, on average, 0.3 minutes longer."

(i) Rate each finding by how well its slope supports it: well, partly, or not at all. Give a one-sentence reason for each rating.

(ii) For each finding rated "partly" or "not at all", state the error in it, and describe a better way to use that variable in a model of trip length.

Answer key

A strong answer: *(i) (a) Not at all: station numbers are IDs, not amounts. (b) Well: temperature and trip length are both amounts, and the finding describes the association between temperature and trip length. (ii) Station 60 is not "more" of anything than station 20, so a slope on station number depends only on the way the IDs were assigned: renumber the stations and the 0.05 changes, although no trip has changed. Start station should enter the model as a categorical variable: a group model that predicts each trip with the mean length of trips from its station (or from its neighborhood) treats each station as its own group, with no order.*

What makes it strong: *A strong answer identifies station number as a label, explains why a slope on a label has no meaning, and replaces the slope with a group model.*

Also right: *"Partly" for (b), with the reason that the slope describes an association, not the effect of warmer weather on a given trip, as long as the answer does not call the number meaningless.*

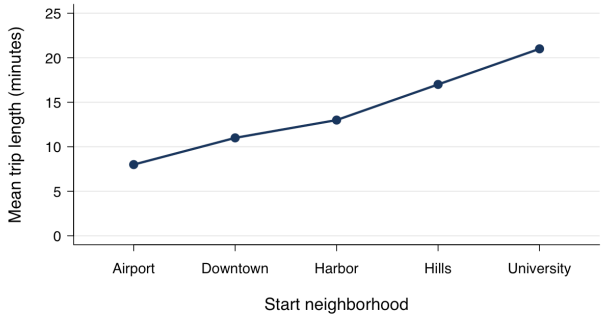
Weaker answers: *Rating (b) "not at all" because "correlation isn't causation": (b) claims only an association. Rating (a) "well" because R produced a slope: R treats any column of numbers as amounts, whether or not the numbers are amounts.*

Rubric (4 points): *(i) 2 points: (a) not at all, (b) well (or partly, as above), 0.5 each for the rating and 0.5 each for the reason. (ii) 2 points: station numbers are labels, so the slope depends on the assignment of the IDs (1); use station or neighborhood as a categorical variable, in a group model (1).*

Key idea: *A slope treats a variable's values as amounts. Before fitting a slope, ask whether a larger value means more of some quantity; a variable whose values are labels belongs in a group model.*

6.2 The neighborhood chart

The analyst made this chart of the mean trip length for trips starting in each of five neighborhoods, and a reader concludes, "Trips get longer as you move along the chart from Airport to University."



(i) Is the variable on the horizontal axis categorical or quantitative? Do its values have a natural order?

(ii) How well do these data support the reader's conclusion: well, partly, or not at all? Explain the reasoning behind your rating.

(iii) Describe the chart you would make to compare the five neighborhoods' mean trip lengths, and explain the reasoning behind your choice.

Answer key

A strong answer: *(i) Start neighborhood, a categorical variable whose categories have no natural order; here the categories are listed alphabetically. (ii) Not at all. The line joining the points suggests a quantity changing steadily from one end of the axis to the other, a trend, but "moving along the chart from Airport to University" is not a direction in the data: list the neighborhoods in another order and the "trend" changes or disappears, although no trip has changed. (iii) A bar chart, or the five means as separate points with no connecting line, one per neighborhood: neither joins the categories as if they lay on a scale. Ordering the bars from the shortest mean to the longest is acceptable, because that order comes from the means themselves, not from a scale on the axis.*

What makes it strong: *A strong answer names the axis variable as an unordered categorical variable, explains that the line imposes an order, and replaces the line with separate bars or points.*

Also right: *"The line also suggests values between two neighborhoods, which have no meaning." Suggesting a chart of every trip's length by neighborhood (for example, side-by-side boxplots), which also shows the variation in trip length within each neighborhood.*

Weaker answers: *Rating the conclusion "well" because the line rises from left to right reads the alphabetical order as a scale. "The line should be straight" or "the axis should start at 5" do not identify the problem. "Use a histogram" confuses the comparison of five groups' means with the distribution of one quantitative variable.*

Rubric (4 points): *(i) 1 point: categorical, with no natural order (1); "categorical" alone: 0.5. (ii) 1.5 points: "not at all" (0.5), because the trend depends only on the order in which the categories were listed (1). "Partly", with that reason: 1 of 1.5. (iii) 1.5 points: bars, or unconnected points, one per neighborhood (1), with the reason that they do not imply an order (0.5). "A histogram": 0 for (iii).*

Key idea: *A line on a graph claims that the horizontal axis is a scale with an order. For unordered categories, compare the groups with bars or separate points.*

6.3 Rain and riding

Over the past 90 days, the analyst compares the days with rain in the forecast with the days without rain in the forecast. The mean number of rides per day was 5% lower on rain-forecast days. When the analyst shuffled the "rain forecast" labels among the days many times, a difference of at least 5% in either direction came up in about 40 of every 100 shuffles. Which is the most accurate summary for the operations team? Choose one, and give your reason.

A. The forecast does not affect ridership, because a difference of this size is nothing more than chance.
B. The forecast lowers ridership by about 5%, since that is the difference observed in the data from these past 90 days.
C. Rides were a little lower, but chance often produces a difference of this size, so the data are not yet decisive.
D. There is about a 40-in-100 chance that the forecast has no effect at all on the number of rides.

Answer key

Answer: *C.*

Why: *A difference of at least 5% came up in about 40 of every 100 shuffles, so chance alone often produces a difference of this size: the result is weak evidence of an effect of the forecast. It is not evidence of no effect: the data still point to fewer rides on rain-forecast days, and a real drop of 5% or more could easily produce the observed result. Data from more days could settle the question.*

Why not the others:

- *A turns "chance could explain it" into "it is chance". A result that chance often produces cannot rule out an effect.*
- *B reads the estimate as established. With a difference that chance often produces, the true effect could be zero, 5% or larger.*
- *D reads the 40 in 100 as the chance that there is no effect. It is the frequency with which chance alone, if the forecast had no effect, produces a difference of at least this size.*

Rubric (4 points): *C (1 point). Reason (3 points): chance often produces a 5% difference, so the evidence is weak (1); weak evidence is not evidence of no effect (1); the estimate still points to fewer rides, and more data could settle the question (1). A with "40 in 100 is common": 0.5 of 3.*

Key idea: *"No effect shown" is not "shown to have no effect". When chance alone often produces a result like the one observed, report that the data are not yet decisive, the direction in which they point, and the data that could settle the question.*

The reasoning moves in this round (Questions 3 to 6)

1. Predict from the overall pattern, only within the range of the data, and only for cases like the ones measured (3.1).
2. Know the alternatives to a causal explanation of a pattern, and which designs allow a causal conclusion and are ethically acceptable (3.2, 3.3).
3. Judge a result by its position among the results of chance alone, and separate weak evidence of an effect from evidence of no effect (4.1, 6.3).
4. Expect estimates from larger samples to vary less than estimates from smaller samples (4.2, 5.2), and know that sample size does not remove the bias from a poorly chosen sample (5.1).
5. Expect false positives among many tests, and compare the number of positive results with the number expected from chance alone (4.3).
6. Expect a result selected for its size to be smaller when repeated, and a real effect to be absent from some studies (4.4).
7. Separate the evidence for an effect from the size of the effect (5.3).
8. Check the meaning of a variable's values before modeling or graphing the variable (6.1, 6.2).