Date 11-02-2026

PSYCH 10 · Fall 2026 · Pre-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 what to carry from the question to new situations.

Question 1 · Two keyboard layouts

A typing course randomly assigned 10 of its students to practice for a week on one of two keyboard layouts, 5 per layout, and then measured each student's typing speed in words per minute (wpm). The R output follows.

```
favstats(~ WPM, data = typing)
 min Q1 median Q3 max mean       sd  n missing
  40 44     46 48  72   48 8.944272 10       0
favstats(WPM ~ Layout, data = typing)
    Layout min Q1 median Q3 max mean        sd n missing
1      New  44 46     48 50  72   52 11.401754 5       0
2 Standard  40 42     44 46  48   44  3.162278 5       0
```

The five New-layout students typed 44, 46, 48, 50 and 72 wpm; the five Standard-layout students typed 40, 42, 44, 46 and 48 wpm.

1.1 One fast typist

The New layout's standard deviation, 11.4 wpm, is more than three times the Standard layout's, 3.2 wpm, yet four of the New layout's five students typed within 6 wpm of one another (44, 46, 48 and 50). The deviations of the five New-layout students from the layout's mean of 52 are −8, −6, −4, −2 and +20.

(i) Which one student accounts for most of the New layout's SS, the sum of squared deviations from the mean? Give that student's typing speed in wpm. Say roughly what share of SS that student accounts for: about half, about three-quarters, or nearly all. Explain how you can tell; you do not need to compute SS exactly.

(ii) Suppose spread were measured with SAD, the sum of absolute deviations, instead. Would that student's share of SAD be larger or smaller than the same student's share of SS? Explain why, in terms of what squaring does to a large deviation.

Answer key

Answer: *(i) The student who typed 72 wpm, with about three-quarters of SS. That student's deviation is +20, so the squared deviation is $20^2 = 400$. The other four squared deviations (64, 36, 16 and 4) add to 120, so SS is 400 + 120 = 520, and 400 is a little more than three times 120: the one student accounts for about three-quarters of SS. (ii) Smaller. With absolute deviations, the same student contributes 20 against the other four students' 8 + 6 + 4 + 2 = 20, so half of SAD rather than three-quarters of SS. Squaring makes a large deviation count far more than a small deviation: a deviation of 20 is 10 times a deviation of 2, but the square of 20 is 100 times the square of 2.*

Also right: *Computing the share exactly (400 ÷ 520, about 77%) is fine but not needed. Giving the deviation (+20) instead of the speed, if the student is clearly identified, earns the identification point.*

Common mistakes: *Judging the share from the deviations rather than from their squares (20 of 40 is half, which is the SAD share, not the SS share). Saying that the share of SAD would be larger "because SAD has no squaring to shrink it". Comparing the 72's 400 in SS with its 20 in SAD directly: the two totals are on different scales, so compare shares.*

Rubric (4 points): *(i) 2 points: the 72-wpm typist (0.5); "about three-quarters" (0.5); the reason, 400 against the other four squared deviations' 120 (1). "About half", from the deviations rather than the squares: 0.5 of 2. (ii) 2 points: "smaller" (0.5); 20 of 40 is half, against three-quarters of SS (0.5); squaring magnifies a large deviation far more than a small deviation (1). A computed share neither adds to nor subtracts from the score.*

The idea to carry with you: *Squared error lets one unusual value dominate a total. Keep that in mind whenever you read an SS, a variance or a standard deviation.*

1.2 Predicting with the median

A classmate says, "The 72 pulls the New layout's mean up, so a better model would predict each layout's *median* instead of its mean."

(i) For the New layout's five students, would predicting the median speed (48) for every student give a larger or a smaller total squared error than predicting the mean speed (52)? Say how you know; you do not need to compute either total.

(ii) In what sense could the classmate's median be the better choice, and in what sense is the mean the better one?

Answer key

Answer: *(i) Larger. For any set of values, the mean is the single number with the smallest total of squared errors, whether or not the set contains an unusual value, so any other prediction, the median included, gives a larger total (here 600 against 520). A second route: the 72 is 24 from the median but 20 from the mean, so the 72's squared error rises from 400 to 576, a rise of 176, while the other four students' squared errors fall from 120 to 24, a fall of only 96. (ii) The 72 has less pull on the median than on the mean, so the median may give a better description of a typical New-layout student. By squared error, the measure used by the models in this course, the mean is the better prediction; the median is the better prediction only by absolute error (a total of 32 against 40 for these five students).*

Also right: *Naming absolute error, by which the median gives the smaller total (32 against 40), as the sense in which the classmate is right.*

Common mistakes: *"Smaller, because the median is less affected by the 72": the median's resistance to an unusual value is a matter of description, not of the total of squared errors. Reasoning from absolute error and concluding that the median is also better by squared error.*

Rubric (4 points): *(i) 2 points: "larger" (0.5); the mean minimizes the total of squared errors, or the 72's squared error rises by 176 while the other four fall by only 96 (1.5). "Smaller, because the median is less affected": 0. (ii) 2 points: the median as a better description of a typical student, since the 72 has less pull on the median (1); the mean as the better prediction by squared error, the course's measure, with the median better only by absolute error (1).*

The idea to carry with you: *Which prediction is "best" depends on the measure of error. With squared error, the measure used by the models in this course, the mean is always the best single prediction.*

1.3 All ten students

The sum of squared deviations is 720 for all 10 students around their overall mean (48) and 520 for the New layout's five students around their mean (52). A classmate concludes, "So the 10 students' speeds are more spread out than the New layout's five."

(i) How well justified is the classmate's conclusion: well, partly, or not at all? Use the standard deviation of all 10 speeds and the standard deviation of the New layout's five speeds, both shown in the output.

(ii) Explain how SS can be larger for the 10 students while their standard deviation is smaller.

Answer key

Answer: *(i) Not at all. The New layout's five speeds are the more spread out: their standard deviation is 11.4 wpm, against 8.9 wpm for all 10 speeds. "Partly" earns credit only with the reason that SS for the 10 is indeed larger, which is a fact about totals, not about spread. (ii) SS is a* total*, so it grows with the number of cases: the 10 students contribute 10 squared deviations, the 5 students only 5. The standard deviation divides SS by n — 1 before taking the square root (720 ÷ 9 = 80 against 520 ÷ 4 = 130), which makes it a per-case measure, so it compares spread fairly across groups of different sizes.*

Also right: *A further correct point, not required: the Standard layout's five speeds (40 to 48) lie close to the overall mean of 48 and add little to SS, which lowers the per-student spread of all 10 students.*

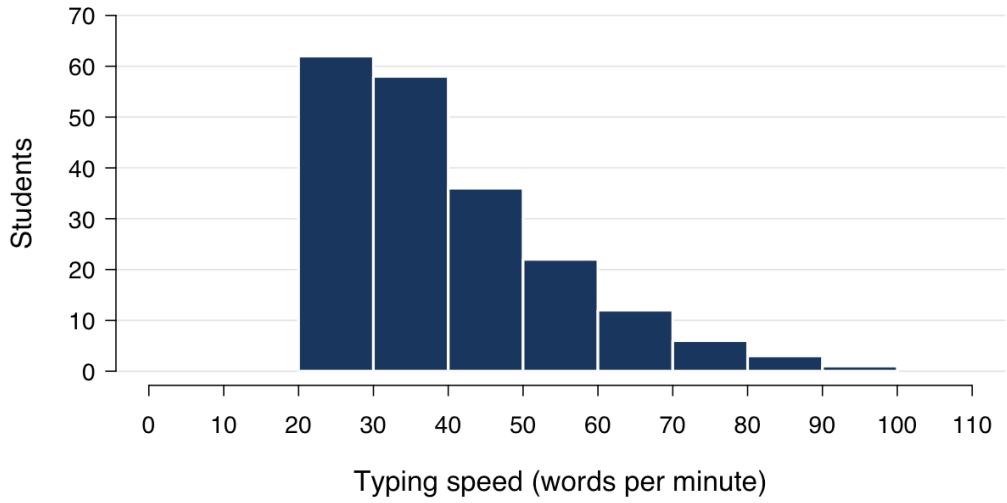
Common mistakes: *"Well justified, because their SS is larger", which compares totals over different numbers of cases.*

Rubric (4 points): *(i) 1.5 points: "not at all" (0.5), with the two standard deviations, 11.4 against 8.9 (1); "partly" with the reason that SS for the 10 is larger as a total: 1. (ii) 2.5 points: SS is a total that grows with the number of cases (1.5); the standard deviation divides by n — 1 and so compares spread per case (1). "SS gets bigger with more students" alone: 1.5.*

The idea to carry with you: *Adding cases can only add squared deviations, so SS grows with the data. To compare spread, use the variance or the standard deviation, which divide by n — 1.*

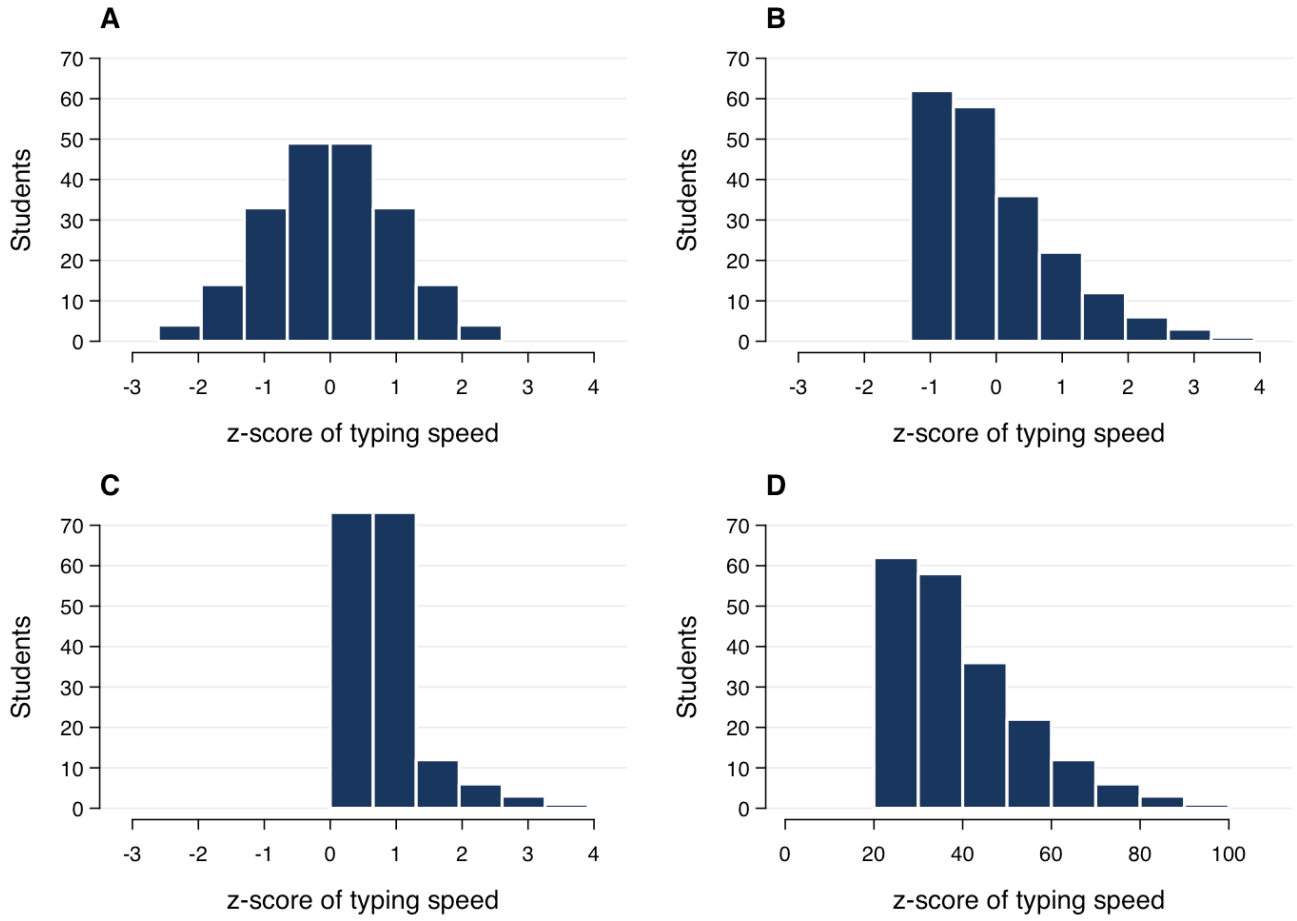
1.4 The whole typing course

All 200 students in the course had their speed measured at the start of term. Their speeds have a mean of 40 wpm and a standard deviation of 15.3 wpm. Here is a histogram showing the distribution of typing speeds (words per minute) in this sample.



(i) The slowest student typed about 20 wpm and the fastest about 95 wpm. About how many standard deviations from the mean is each? A rough number is enough (for example, "about 1" or "between 3 and 4"). Which of the two speeds, 20 wpm or 95 wpm, is the more unusual?

(ii) Suppose all 200 speeds were converted to z-scores. Which of the four histograms below shows what you expect the histogram of the 200 z-scores to look like? Choose one, and give your reason.



(iii) In a normal (bell-shaped) distribution, about 2.5% of values lie more than two standard deviations above the mean, and about

2.5% lie more than two standard deviations below the mean. For these speeds, two standard deviations above the mean is about 70 wpm, and two below is about 10 wpm. Use the histogram to judge whether each of those two percentages holds for these 200 students: is about 2.5% of the course above 70 wpm, and is about 2.5% below 10 wpm? Explain any difference.

Answer key

Answer: *(i) The slowest is about 1 standard deviation below the mean: 20 wpm is 20 wpm below 40, and one standard deviation is about 15 (the z-score is about —1.3). The fastest is between 3 and 4 standard deviations above the mean: 95 wpm is 55 wpm above 40 (about 3.6 standard deviations). The fastest is by far the more unusual of the two. (ii) B. z-scoring subtracts the mean from every speed and divides by the standard deviation. That relabels the axis (the slowest at about —1.3, the fastest at about 3.6) without moving the bars relative to one another, so the z-scores have exactly the same right-skewed shape as the speeds, centered at 0 with a standard deviation of 1. Why not the others: A is bell-shaped, but standardizing does not change a distribution's shape; C puts every z-score at or above 0, but every speed below the mean has a negative z-score; D is the original histogram unchanged, but the z-scores are centered at 0 with a standard deviation of 1, not at 40 wpm with a standard deviation of 15. A check: the share of z-scores above 2 is the share of speeds above 70 wpm, and (iii) shows that share to be about twice the normal distribution's 2.5%, so the z-scores are not bell-shaped. (iii) Above 70 wpm, the bars hold about 10 students (the 70 to 80 bar reads as about 6, the 80 to 90 bar as about 3, the 90 to 100 bar as 1), about 5% of the 200, twice the normal distribution's 2.5%; a reading anywhere from 8 to 12 students, 4% to 6%, is fine. No student is below 10 wpm (the slowest is at about 20), so that share is 0% rather than 2.5%. The percentages describe a normal distribution; these speeds are skewed to the right, with a long tail of fast typists and a floor on the slow side.*

Common mistakes: *For (i), reporting the distances in wpm (20 and 55) without dividing by the standard deviation. For (ii), choosing A "because z-scores go with the normal distribution" (standardizing does not reshape the data), or D because "standardizing just relabels" (the relabeling moves the center to 0 and the spread to 1, and D's axis still runs from 20 to 100). For (iii), judging only the upper tail.*

Rubric (5 points): *(i) 1 point: about 1 standard deviation below and 3 to 4 above (0.5); the fastest is the more unusual (0.5). Distances in wpm only: 0. (ii) 2 points: B (0.5); standardizing relabels the axis without changing the shape, center 0 and standard deviation 1 (1.5). A: 0. D with "it just relabels": 0.25. (iii) 2 points: the upper share about 4% to 6%, above 2.5% (0.5); the lower share 0% (0.5); the percentages describe a normal distribution and these speeds are right-skewed with a floor (1).*

The idea to carry with you: *z-scores keep a distribution's shape, so the normal distribution's percentages hold only when the data themselves are bell-shaped.*

Question 2 · Checking an AI assistant's ANOVA table

A company tested two versions of its checkout page. It randomly assigned 24 shoppers, 12 to each version, and timed how many seconds each shopper took to finish checking out. The company asked an AI assistant to analyze the data. Instead of showing R's output, the AI assistant wrote out the table and the summary below on its own. AI assistants can report numbers that look right but are wrong.

```
Model: Seconds ~ Page
             SS  df      MS      F     PRE
Model    44.000   1  44.000  4.000  0.6667
Error    88.000  23   4.000
Total   132.000  23
```

"*Shoppers using the new page finished checking out faster on average than shoppers using the old page, and the page explains two-thirds of the variation in checkout time (PRE = .67).*"

2.1 Which numbers don't add up?

Exactly three of the numbers in the table cannot be right, given the other numbers and the 24 shoppers. Use the relationships that tie the entries of an ANOVA table together (df, SS, MS, F and PRE) to find the three. For each, name the relationship it breaks and say what the number should be.

Answer key

The three wrong numbers:

- ***df Error = 23.*** *With 24 shoppers, df Total = 24 — 1 = 23, and a two-group model uses df Model = 2 — 1 = 1, so df Error = 23 — 1 = 22. The table's own MS Error confirms the correction: 88 ÷ 22 = 4, the value shown, whereas 88 ÷ 23 ≈ 3.8. (Error and Total cannot both be 23.)*
- ***PRE = .6667.*** *PRE = SS Model ÷ SS Total = 44 ÷ 132 = 1/3, about .33. The AI assistant divided SS Error by SS Total, which gives the unexplained share of the variation rather than the explained share.*
- ***F = 4.000.*** *F = MS Model ÷ MS Error = 44 ÷ 4 = 11. The reported 4 is the MS Error, not F.*

Consistent numbers: *SS Model + SS Error = SS Total (44 + 88 = 132). MS Model = 44 ÷ 1 = 44. MS Error = 88 ÷ 22 = 4.*

What makes it strong: *A strong answer checks each number against the others with a relationship that ties the table's entries together, rather than judging whether a number looks plausible.*

Weaker answers: *Flagging SS Total or MS Model, which are consistent. Saying "F = 4 looks small" without naming F = MS Model ÷ MS Error, the relationship that gives 11. Flagging MS Error (88 ÷ 23 ≈ 3.8) after correcting df Error to 22 is self-contradictory: with df Error = 22, MS Error = 4 is consistent. "Exactly three" is a check: a fourth flag means that one of the four is wrong.*

Rubric (6 points): *2 points for each wrong number (df Error, 22; PRE, 1/3; F, 11): name the number (0.5), name the broken relationship (1), give the corrected value (0.5). Flagging a consistent number: —1, to a floor of 0; the self-contradictory MS Error flag after a df correction counts as a wrong flag.*

The idea to carry with you: *The entries of an ANOVA table are tied together: df Total is one less than the number of data points, SS Model and SS Error add to SS Total and df Model and df Error add to df Total, MS is SS ÷ df, F is MS Model ÷ MS Error, and PRE is SS Model ÷ SS Total. When software, or an AI assistant, reports a table, those ties let you check the table.*

2.2 Time of day instead

The 24 shoppers came at four times of day, 6 at each. A designer fits `lm(Seconds ~ TimeOfDay)` to the same data and gets PRE = .40. The designer says, "Time of day explains 40% of the variation in checkout time, so it is the model to use." Compare it with the Page model, using the Page model's corrected PRE and F from 2.1.

(i) Would the time-of-day model's F be larger or smaller than the Page model's corrected F? Explain, using each model's PRE and the number of parameters each model uses (its df Model).

(ii) Which model gives the more convincing evidence that it explains real differences in checkout time, and why?

Answer key

Answer: *(i) Smaller (about 4.4). The time-of-day model's PRE of .40 is spread over 3 df Model, about .13 per parameter, against .33 for the Page model's single parameter. F compares the error explained per model df with the error left per error df, so the Page model's F is larger, 11 against about 4.4. A model with more groups can always explain more error, even by chance alone, so a model's PRE has to be judged against its number of parameters. (ii) The Page model. With one parameter it explains nearly as much error (one-third) as the time-of-day model explains with three (.40), so its F is more than twice as large, and chance alone is a far less likely explanation of its result.*

Also right: *For (ii), the point that only the page was randomly assigned: the shoppers came at times of day they chose, so even a real time-of-day difference could reflect the kinds of shoppers who come at each time, while a page difference can be attributed to the page. Follow-through from 2.1: a student whose corrected PRE or F in 2.1 differs from 1/3 and 11 earns full credit here for reasoning that compares the two models' PRE per df Model, and their Fs, consistently with the student's own values.*

Common mistakes: *"Larger, because its PRE is higher." Counting the time-of-day model's parameters as 4 rather than 3.*

Rubric (4 points): *(i) 2.5 points: "smaller" (0.5); the PRE of .40 is spread over 3 df Model against .33 over 1, so the Page model explains more per parameter, and more groups always raise PRE even by chance (2). "Larger, because its PRE is higher": 0. (ii) 1.5 points: the Page model (0.5), with the reason that its PRE per parameter, and so its F, is larger, so chance alone is a less likely explanation of its result, or that only the page was randomly assigned (1).*

The idea to carry with you: *Extra parameters always raise PRE by some amount. F weighs the PRE against the number of parameters the model used, so the simpler of two models can be the more convincing model.*

Question 3 · The city's animal shelters

A city runs four animal shelters. The table gives the number of dogs each shelter placed in new homes in 2026 and the mean number of days those dogs waited before adoption.

Shelter	Dogs adopted	Mean wait (days)
North	200	19
South	50	40
East	30	50
West	20	70

3.1 The citywide mean

An analyst adds up the four means and divides by 4: (19 + 40 + 50 + 70) ÷ 4 ≈ 45. The report says, "The mean wait of an adopted dog in our city was 45 days."

(i) Each row of the table is one shelter. In the analyst's 45, does each shelter count equally, or does each dog count equally?

(ii) North placed 200 of the 300 dogs, and those 200 dogs' mean wait was 19 days. Is the mean wait per dog closer to 19 days or to 45 days? Explain how you can tell; you do not need to compute the per-dog mean exactly.

(iii) How well justified is the report's sentence? Rate it as well justified, partly justified, or not justified, and give the specific reason for your rating in one or two sentences.

Answer key

Answer: *(i) Each shelter. The 45 is the mean of four shelter means, so West's 20 dogs carry the same weight as North's 200. (ii) Closer to 19 days. North's 200 dogs are two-thirds of all 300, and their mean wait is 19 days; the other 100 dogs, with shelter means of 40, 50 and 70 days, raise the per-dog mean only part of the way toward 45. (The per-dog mean is 29 days: 8,700 dog-days of waiting divided by 300 dogs.) (iii) Not justified. The sentence is about dogs, but the 45 is a mean over shelters, in which West's 20 dogs count as much as North's 200; the mean wait per dog is about 29 days, so the sentence overstates a typical dog's wait by more than half. "Partly justified" earns credit with the reason that 45 correctly answers a different question, the mean of the four shelter means.*

Why: *A mean over dogs counts each shelter's mean once for each of its dogs. A mean of the four rows counts each shelter once, whatever its size.*

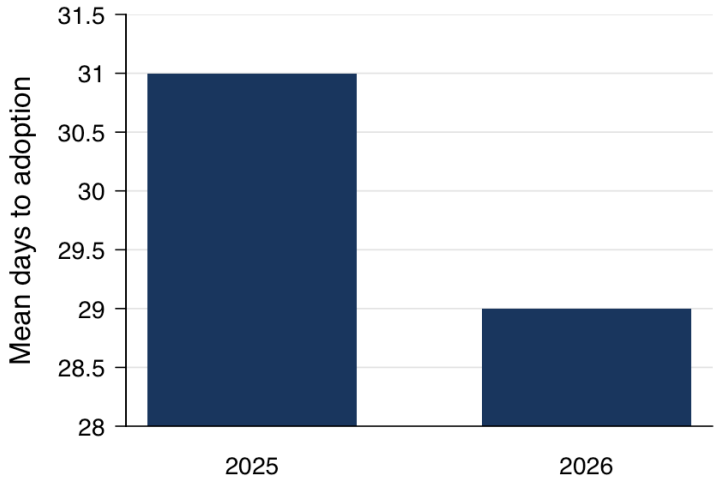
Weaker answers: *"Closer to 45, because 45 is the mean of the four means" repeats the analyst's error. Adding the four shelter means (179 days) as if they were days waited, then dividing by 300, gives less than one day per dog; a result that small is a signal to recheck the arithmetic. Dividing 179 by 4 instead returns to the analyst's 45.*

Rubric (4 points): *(i) 0.5 point: "each shelter". (ii) 1.5 points: "19 days" (0.5) and the reason that two-thirds of the dogs are North's, whose mean wait is 19 days (1). (iii) 2 points: "not justified" with the reason that the 45 is a mean over shelters while the sentence is about dogs (2); "partly justified" with the reason that 45 answers a different question (1); "well justified" (0). An exact per-dog mean neither adds to nor subtracts from the score.*

The idea to carry with you: *Before summarizing a table, identify the cases in its rows, here shelters rather than dogs. A mean over shelters answers a question about shelters. A question about dogs needs a mean over dogs.*

3.2 The director's slide

The shelters' director shows this chart at a city council meeting and says, "Wait times fell by about two-thirds in one year."



(i) What feature of the chart makes the drop look like two-thirds?

(ii) Read the 2025 mean and the 2026 mean off the axis. By about how many days did the mean wait fall, and how does that drop compare with the two-thirds the director claims?

(iii) Describe a chart of the 2025 and 2026 means that would not exaggerate the change.

Answer key

A strong answer: *(i) The vertical axis starts at 28 days rather than at 0, so each bar shows only the part of its mean above 28 days. The 2025 bar is 3 units tall and the 2026 bar 1 unit: a two-thirds drop in bar height for a 2-day drop in the mean. (ii) The mean fell from 31 to 29 days, a drop of 2 days, about 6% of the 2025 mean: a small fraction of the director's two-thirds. (iii) A bar chart whose vertical axis starts at 0, on which the 2026 bar is about 94% as tall as the 2025 bar; or the two means stated in words, with no chart.*

What makes it strong: *It names the feature of the chart that exaggerates the change, reads the real change off the axis and compares that change with the claim, and describes a chart whose bars start at 0.*

Weaker answers: *"The chart needs more years" or "it should be a line graph" do not say what in this chart exaggerates the drop.*

Rubric (3 points): *(i) 1 point: the vertical axis starts at 28 rather than 0 (0.5), so the bars show only the part of each mean above 28 and stand in a 3-to-1 ratio (0.5). (ii) 1 point: 31 and 29, a drop of about 2 days (0.5); a small fraction of the claimed two-thirds (0.5). (iii) 1 point: a chart whose axis starts at 0, or the two means stated in words (0.5); why that chart shows the change fairly, for example the 2026 bar about 94% as tall (0.5). "More years" or "a line graph": 0 for (iii).*

The idea to carry with you: *Readers compare bars by their lengths. When the axis does not start at 0, bar lengths exaggerate the differences between the values, so read the axis before the bars.*

Question 4 · Who uses the library?

A city wants to estimate what share of its adult residents used a public library in the past year.

4.1 The library's list

The library chose 500 people at random from its list of library-card holders and asked whether they had used a library in the past year. Of the 500, 58% said yes.

(i) As an estimate of the share of all adult residents who used a library in the past year, is 58% likely to be too high, too low, or about right? Explain how the way the sample was chosen affects the estimate.

(ii) How much would choosing 5,000 card holders at random instead improve the estimate of that share: a lot, a little, or not at all? Explain, and say how the sample would have to be chosen for the estimate to describe all adult residents.

Answer key

A strong answer: *(i) Probably too high. The sample was chosen at random, but from the wrong list: only card holders could be chosen, and residents without a card are less likely to have used a library, so the sample's share overstates the share among all adult residents. (ii) Not at all, as an estimate of the share among all adult residents. A random sample of 5,000 card holders would estimate a different quantity, the share among card holders, more precisely; it would still exclude every resident without a card, so its bias is unchanged. For the estimate to describe all adult residents, the sample has to be drawn from a list of all adult residents.*

What makes it strong: *It recognizes that random choice from an incomplete list does not remove the bias, and that a larger sample from the same list has the same bias.*

Also right: *For (ii), "a little", with the reason that the larger sample is more precise about card holders while the estimate of the resident share keeps its bias; "a more precise estimate of the wrong quantity" or "more precisely wrong". For the list, any source that covers every adult resident: addresses, voter rolls, random-digit dialing.*

Weaker answers: *"About right, because the sample is random" overlooks the list the sample came from. "A lot, because the sample is bigger" confuses precision with lack of bias. "A bigger sample is more accurate, but it would still be only card holders" is partly right: it names the list as the problem but does not separate precision, which the larger sample improves, from bias, which the larger sample leaves unchanged.*

Rubric (5 points): *(i) 2.5 points: "too high" (0.5); the list holds only card holders, who use libraries more, and random selection within the list does not correct the list (2). (ii) 2.5 points: "not at all" (0.5); a larger sample from the same list improves precision about card holders and leaves the bias unchanged (1.5); the sample must be drawn from a list of all adult residents (0.5). "A little" with the precision reason: 0.25 for the rating, full credit for the reasons. "A lot, because bigger": 0 for the rating and its reason; 0.5 if the correct list is still named.*

The idea to carry with you: *A random sample represents only the list it was drawn from. Check that the list matches the population of interest.*

4.2 How far off?

The city then surveyed a random sample of 400 of all adult residents. The 400 residents' mean number of library visits in the past year was 9, with a standard deviation of 12 visits (many visited 0 times; a few visited more than 50 times). A planner says, "With a standard deviation of 12, our sample's mean of 9 could easily be off by 12 visits." Which response is best? Choose one, and give your reason.

A. Agree: a sample's mean is usually off by about one standard deviation, which here is 12 visits.
B. Agree: many residents visited 0 times, so the data are too uneven for their mean to be a useful estimate.
C. Disagree: a mean of 400 residents varies far less than single residents vary, so the sample's mean is probably within a visit of the city's.
D. Disagree: because the sample was chosen at random, its mean equals the mean for all adult residents exactly.

Answer key

Answer: *C.*

Why: *The standard deviation of 12 describes the spread among* individual residents. *The mean of 400 residents varies much less from sample to sample, because within any sample of 400 the unusually high values are largely offset by the unusually low values. Different random samples of 400 would give means that differ from one another, and from the city's mean, by about half a visit to one visit.*

Why not the others:

- *A uses the spread of individual values as if it were the spread of sample means.*
- *B is right that the data are uneven, but the mean of 400 uneven values is still a stable estimate.*
- *D claims too much for random sampling: a random sample's mean is close to, but not equal to, the mean for all residents.*

Weaker answers: *C with only "400 is a big sample, so the mean should be close" has the right choice and the germ of the reason, but not the distinction between the spread of individual values and the spread of sample means. A rough size for the spread of sample means, 12 ÷ $\sqrt{400} \approx 0.6$ visits, is a welcome addition and is not needed to choose C.*

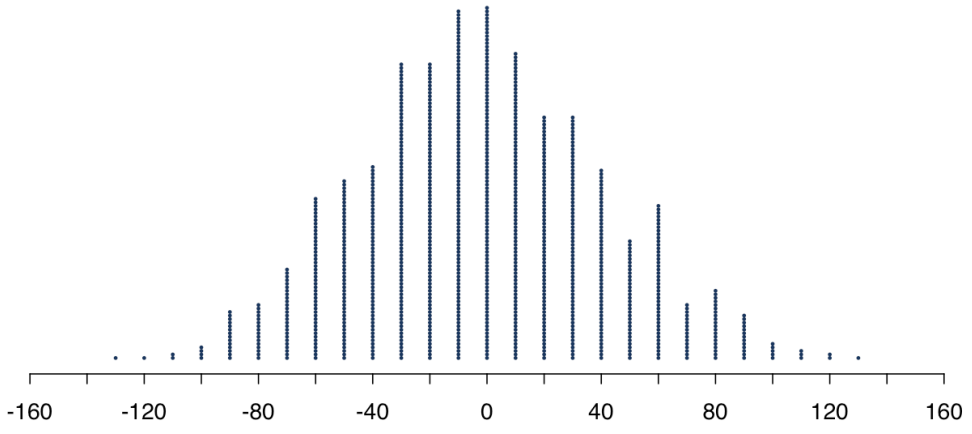
Rubric (3 points): *C (1 point). Reason (2 points): the standard deviation of 12 describes the spread among individual residents (1); the mean of 400 varies far less because high values are offset by low values (1). "400 is a big sample, so the mean is close" alone: 0.5 of 2.*

The idea to carry with you: *Individual values vary widely; a mean of many values varies within a far narrower range. Do not judge the precision of a mean by the standard deviation of the individual values.*

Question 5 · Calorie counts on the menu

A restaurant randomly assigned each of 60 diners to one of two menus: 30 received a menu with the calories listed next to each dish, and 30 received the same menu without calories. The mean calories ordered were 90 lower with the calorie menu (796 against 886).

To judge this result, the researchers wrote each diner's calories ordered on a card, shuffled the 60 cards, dealt them into two piles of 30, called one pile "pretend calorie menu" and the other "pretend plain menu", and computed the difference between the two piles' mean calories (pretend calorie menu minus pretend plain menu). The researchers repeated this shuffle-and-deal 1,000 times. The plot shows the 1,000 differences, one dot per repetition.



5.1 What the shuffled differences show

Which statement best describes the 1,000 shuffled differences? Choose one, and give your reason.

A. The differences that chance alone would produce if the menu had no effect on diners' orders.
B. The differences that would be found if the study were repeated with 60 new diners.
C. The amount by which each diner's order would have changed with the other menu.
D. The differences that chance alone would produce if calorie labels cut orders by 90 calories.

Answer key

Answer: *A.*

Why: *Shuffling the cards discards the record of each diner's menu. Each diner's calories stay the same; only the pretend labels "calorie menu" and "plain menu" are dealt out at random, whatever menu each diner actually had. Such a reassignment is exactly the situation in which the menu has no effect: every diner would have ordered the same calories with either menu, and the difference between the two groups would come only from which diners happened to be dealt into which pile.*

Why not the others:

- *B describes new diners. The 1,000 repetitions are repetitions of the shuffle, not of the study: nothing new is measured, and every repetition reuses these 60 diners' orders.*
- *C is a claim about each individual diner, which the study cannot observe: each diner saw only one menu.*
- *D describes a situation in which labels have an effect of 90 calories. The shuffled differences describe the situation in which labels have no effect, which is the comparison the real result is judged against.*

Rubric (3 points): *A (1 point). Reason (2 points): shuffling breaks the link between menu and calories, so each difference is one that chance alone would produce if the menu had no effect (2). "Because the piles are random" with no mention of the no-effect situation: 0.5 of 2. "It shows the no-effect case" with no link to the shuffling: 1.5 of 2.*

The idea to carry with you: *A simulation like this one shows the differences chance alone would produce if there were no effect. The real result is judged by its position among those chance differences.*

5.2 Ninety calories

The researchers would have been just as interested if the calorie menu had *raised* orders by 90 calories.

(i) Use the plot to estimate roughly how many of the 1,000 shuffled differences are at least 90 calories from 0, in either direction (counting to the nearest 10 is fine), and say roughly what fraction of the 1,000 that is.

(ii) State what that fraction says about the real result: how often chance alone would produce a difference of at least 90 calories if the menu had no effect, and whether the real result is strong or weak evidence that the calorie menu changed diners' orders.

Answer key

A strong answer: *(i) About 46 of the 1,000 shuffled differences (22 + 24), roughly 1 in 22, or about 1 in 20. Left tail: the columns at —90 (about 14), —100 (4), —110 (2), —120 (1) and —130 (1), about 22 in all. Right tail: the columns at 90 (about 13), 100 (5), 110 (3), 120 (2) and 130 (1), about 24 in all. The columns at exactly —90 and 90 count, because the prompt says "at least". Any total from about 36 to 60 that includes both tails and the columns beyond 490 is right; that total is about 1 in 20, or 5%, of the 1,000. A count of about 27 from the two 490 columns alone (14 + 13) omits the more extreme differences, and a count of 20 to 25 from one tail answers a different question. (ii) If the menu had no effect, chance alone would produce a difference of at least 90 calories, in either direction, only about 1 time in 20. The real result is therefore fairly strong evidence that the calorie menu changed diners' orders. "Moderate", or "fairly strong, but only just", are equally good readings of a result that chance produces about 1 time in 20; "weak" and "proves" are not.*

What makes it strong: *It counts both tails, because a difference in either direction would have counted as evidence, and it states how rarely chance alone produces a difference of at least 90 calories.*

Weaker answers: *Counting only the left tail answers a different question. "Only 46 of 1,000, so there is a 95% chance that labels work" mistakes the share of shuffled differences for the probability that the labels have an effect; the simulation gives the first, not the second.*

Rubric (5 points): *(i) 1.5 points: both tails counted (0.5); the 490 columns and every column beyond them included, a total of about 36 to 60 (0.5); the total expressed as a fraction, about 1 in 20 or 4% to 6% (0.5). One tail only: 0.5 of 1.5. The two 490 columns only, about 27: 1 of 1.5. (ii) 3.5 points: the fraction is how often chance alone would produce a difference of at least 90 calories if the menu had no effect (1.5); "fairly strong" or "moderate" evidence, with the reason that chance rarely produces such a difference (2). "A 95% chance that labels work": 0.5 of 3.5. "Weak" or "proves": 0 for the verdict.*

The idea to carry with you: *Judge a result by how often chance alone produces a result at least as extreme, counting every outcome that would have counted as evidence, in either direction. When the groups were formed by random assignment, a difference that chance rarely produces points to the treatment itself rather than to some difference between the groups.*

5.3 Twenty other things

Besides calories, the researchers compared the two groups on 20 other outcomes they had recorded, such as drinks ordered, desserts ordered, minutes at the table, and tip. Only one of the 20 comparisons, the tip, showed a difference of a size that came up in fewer than 5 of every 100 shuffled repetitions: diners with the calorie menu tipped more. A news story says, "Calorie labels make diners tip more!"

(i) How well justified is the news story's conclusion: well, partly, or not at all? Explain, in terms of the way the researchers arrived at the tip result.

(ii) Write one or two sentences giving a fair report of the tip result.

Answer key

A strong answer: *(i) Not at all. If the menu affected none of the 20 other outcomes, each comparison would still have about a 5-in-100 chance of showing a difference of this size, so across 20 comparisons about one such difference (20 × 5/100) is expected from chance alone. One such difference among 20 comparisons is exactly the number expected from chance alone, so the tip result is not evidence that labels affect tips. "Partly" earns credit only with the reason that the tip difference may be real but needs a study of its own, and only if the number of comparisons is named. (ii) "The researchers also compared the two groups on 20 other outcomes. Only one of them, the tip, showed a difference larger than the usual chance difference; about one such difference is expected by chance among 20 comparisons, so the tip result needs a study of its own before it counts as a finding."*

What makes it strong: *It connects the 5-in-100 chance to the number of comparisons and says that the number of comparisons belongs in the report.*

Weaker answers: *"The tip result is significant, so it's real" ignores the other 19 comparisons. "The study is worthless" overreacts: the calorie comparison was the one comparison planned in advance. "Correlation isn't causation" misreads the design: the menus were randomly assigned, so if the tip difference were real, the labels would be its cause; the question is whether the difference is real, given that it was one of 20 comparisons. A report that says only "more research is needed" does not say why; a fair report states the number of outcomes compared.*

Rubric (5 points): *(i) 3 points: "not at all" (0.5); each of the 20 comparisons has about a 5-in-100 chance of a difference this size, so about one is expected from chance alone (2.5). "Partly" with the retest reason and the number of comparisons named: 2. "Correlation isn't causation": 0.5. (ii) 2 points: a report that states the number of comparisons (1.5) and treats the one standout as the expected chance result, a lead for a new study (0.5). "More research is needed" alone: 0.5.*

The idea to carry with you: *Compare many outcomes, and some will look surprising by chance. Report the number of outcomes compared, and treat a lone surprising one as a lead, not a finding.*

5.4 The second restaurant

A second restaurant repeats the study with 30 diners, 15 per menu. The mean calories ordered were 60 lower with the calorie menu. The researchers repeated the shuffling procedure on the new data: a difference of at least 60 calories from 0, in either direction, came up in about 27 of every 100 repetitions. A headline says, "Calorie labels don't work after all: study fails to replicate."

(i) How well does the second study's result agree with the first study's: fully, partly, or not at all? Compare the two differences in mean calories (their direction and size), and say what the 27 in 100 tells you about the second result on its own.

(ii) What does the second study, on its own, allow you to say about the size of the effect of calorie labels, if any? Then say how well the headline's conclusion is supported: well, partly, or not at all.

Answer key

A strong answer: *(i) Fully. Compare the two estimates: both studies found lower mean calories with the calorie menu, 60 lower against 90 lower, the same direction and a similar size, so the second result does not point to a different effect. Compare the two chance rates: 27 in 100 in the second study against about 5 in 100 in the first. With only 15 diners per menu, chance alone produces a difference of 60 calories or more about one time in four, so the second study on its own cannot distinguish its 60-calorie difference from chance. (ii) Little. The second study's data are as consistent with no effect as with an effect of 60 calories, with an effect of 90 like the first study's, or with a larger effect: with 15 diners per menu, the shuffled differences spread so widely that each of these values is plausible. On its own, the second study says almost nothing about the size of the effect. The headline's conclusion is not supported at all: "fails to replicate" here means that the second study was too small to tell an effect from chance, not that it showed no effect. Taken together, the two studies still point to labels lowering orders, by an amount that remains uncertain.*

What makes it strong: *It compares the two estimates rather than the two verdicts, ties the 27 in 100 to the second study's size, and separates "no effect shown" from "shown to have no effect".*

Also right: *"Partly", with the comparison stated (the same direction, a smaller size). Comparing the second study's 27 in 100 with the*

first study's 5 in 100. Using the spread of the first study's shuffled differences as a reference for whether 60 differs from 90 by more than chance (60 is well inside the range chance covered in the first study). "The second study has no evidence of an effect, which is not evidence of no effect."

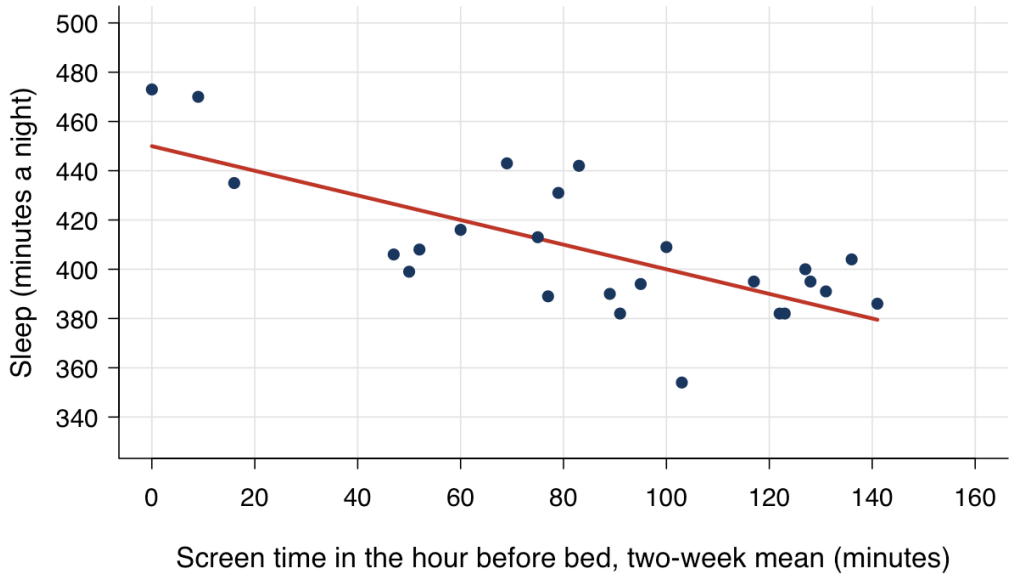
Weaker answers: *"One is significant and one isn't, so they disagree" compares verdicts rather than results. "The second study proves labels don't work" treats a lack of evidence as evidence of no effect. "No, because 60 is still a difference" is right about the direction but does not use the 27 in 100: a difference of 60 is also a common result of chance alone in a study this small.*

Rubric (6 points): *(i) 3 points: "fully" or "partly", with the two estimates compared (same direction, 60 against 90) (1.5); the meaning of the 27 in 100, that with 15 per menu chance alone often produces a 60-calorie difference, so the second study cannot tell its result from chance (1.5). "Not at all" or "they disagree because one is significant": 0.5 if the estimates are compared, 0 otherwise. (ii) 3 points: little can be said about the size, since the data are as consistent with no effect as with 60 or 90 (1.5); the headline is not supported at all, because failing to detect an effect is not showing no effect (1.5). "No, because 60 is still a difference" without the chance point: 0.5.*

The idea to carry with you: *A repeated study gives a somewhat different result, and a smaller study gives a less precise result. Judge a replication by whether its estimate agrees with the original estimate, not by whether both results pass the same threshold; and ask how large an effect a small study could have detected before reading its "failure" as a finding.*

Question 6 · Screens before bed

A sleep researcher asked 25 adults to record, every night for two weeks, the minutes they spent on screens in the hour before bed and the minutes they slept. In the plot, each dot is one adult's two-week means. The red line is the best-fitting straight line: predicted sleep = 450 − 0.5 × screen minutes.



6.1 Reading the line

(i) State in words, with units, what the −0.5 in the line's equation says about these 25 adults.

(ii) One adult's two-week means were 69 minutes of screen time and 443 minutes of sleep. Find this adult's dot in the plot. Is the adult's residual (actual sleep minus the line's predicted sleep) positive or negative, roughly how many minutes is it (reading from the plot is fine), and what does it say about this adult?

Answer key

A strong answer: *(i) Among these 25 adults, each additional minute of screen time before bed is associated with about half a minute less sleep a night; for example, an adult with 10 more minutes of screen time is predicted to sleep about 5 minutes less per night. The −0.5 is a slope: a difference in predicted sleep between adults, not the effect of changing one adult's screen time. (ii) Positive. At 69 minutes of screen time the line predicts about 415 minutes of sleep (450 − 0.5 × 69 = 415.5), and the dot sits about 28 minutes above the line (443 − 415.5 = 27.5). This adult slept about half an hour a night more than the line predicts for that much screen time. Read from the plot, any value from about 20 to 35 minutes above the line is right; computing gives 27.5, also right. The sign must be positive.*

What makes it strong: *(i) states the slope in the variables' units, as a difference between adults rather than as a change within one adult. (ii) locates the dot, reads the sign of the residual from the dot's position above the line, and gives a rough size.*

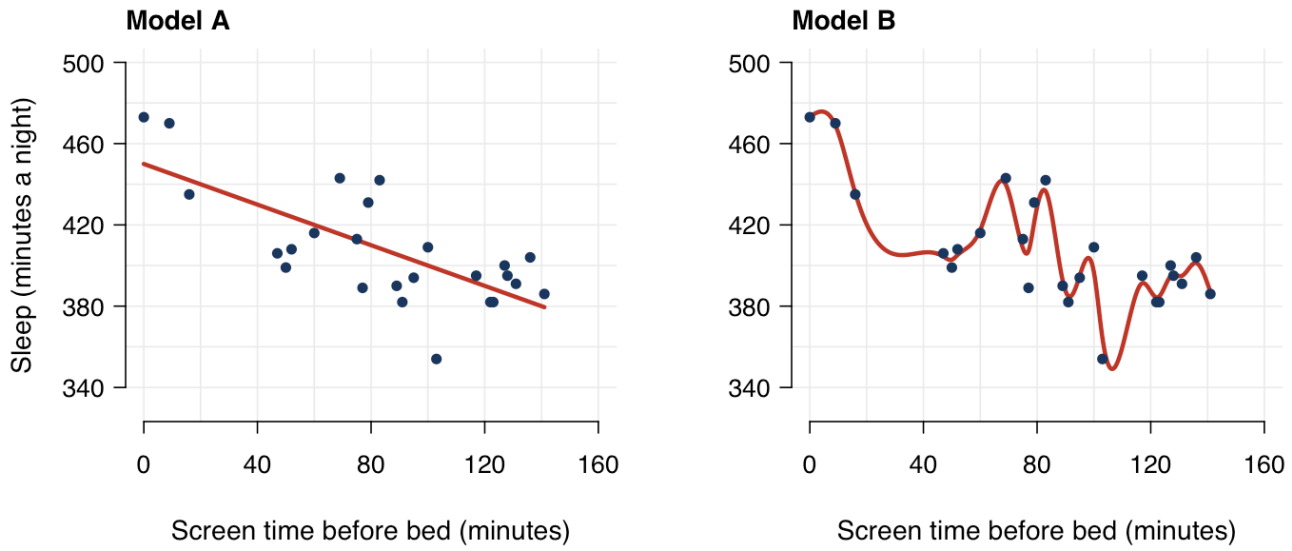
Weaker answers: *"Screens reduce sleep by 0.5" states the slope as a causal effect and omits the units. Taking predicted minus actual gives the right size with the wrong sign, and so makes this adult's sleep look shorter than the prediction.*

Rubric (3 points): *(i) 2 points: about half a minute less sleep per additional minute of screen time, with units (1); stated as an association or prediction among these adults, not as the effect of changing one adult's screen time (1). (ii) 1 point: positive (0.25); roughly 20 to 35 minutes, or 27.5 if computed (0.25); the adult slept about half an hour more than the line predicts for that screen time (0.5). Wrong sign: 0 for the sign, and 0.25 for the meaning if the sign rule is stated correctly.*

The idea to carry with you: *The slope is the change in the predicted value for each one-unit increase in the explanatory variable. A residual is the distance of one case from its prediction, with a sign that says whether the case lies above or below the line.*

6.2 A curve through every point

A classmate fits Model B, a curve that bends to pass near almost every point. Its total squared error (the sum, over the 25 adults, of each adult's squared residual) is 863, against 9,167 for the straight line, Model A.



A new adult has a two-week mean of 105 minutes of screen time before bed.

(i) Read each model's predicted sleep for this adult from the figure, to the nearest 10 minutes.

(ii) Say which prediction you trust more for this new adult, and explain why. Your explanation should say what the two models' total squared errors, 863 and 9,167, tell you about predictions for a new adult.

Answer key

A strong answer: *(i) Model A predicts about 400 minutes ($450 - 0.5 \times 105 = 397.5$). Model B predicts about 350 minutes, because near 105 minutes the curve dips to pass near one adult with unusually short sleep. (ii) Model A's prediction. Model B's small total squared error measures its fit to these 25 adults: the curve follows each adult's chance departure from the trend, and those departures will not recur in a new adult. The dip near 105 minutes reflects one person, not a pattern. The straight line follows the overall trend, so its prediction is the better one for a new adult. The two squared errors say which model fits the old data more closely, not which model gives the better prediction for a new case.*

What makes it strong: *It reads both predictions from the figure, traces Model B's prediction to one unusual adult, and judges the models by their predictions for new adults rather than by their fit to these 25 adults.*

Weaker answers: *"Model B, because its error is smaller" judges the model only on the data it was built from. "Smaller error means a better model, like the mean" carries over a rule from Chapter 5 that holds when single-number predictions are compared on the same data; the rule does not hold when one model has far more freedom to bend, because the extra freedom fits the chance departures of these 25 adults, and a new adult brings new departures. "Model A, because straight lines are simpler" picks the right model without saying why the simpler model gives the better prediction in this case.*

Rubric (5 points): *(i) 1 point: about 400 for Model A and about 350 for Model B (0.5 each; 390 to 400 and 340 to 360 accepted). (ii) 4 points: Model A (0.5); the two squared errors measure fit to these 25 adults, not accuracy for a new adult (2); Model B's flexibility follows chance departures, and the dip near 105 is one adult (1.5). "Model A, because straight lines are simpler": 2 of 4. "Model B, because its error is smaller": 0.5 of 4, for naming the errors.*

The idea to carry with you: *A model that bends to follow every point fits the chance variation in the data as well as the pattern. Predict from the overall trend.*

6.3 The headline

A news story about the study is headlined, "Screens before bed steal your sleep."

(i) The headline assumes that the causation runs in one direction. State that direction in one sentence.

(ii) Describe one specific process by which the pattern in the plot could arise even if screens do not shorten sleep.

(iii) What does the possibility you described imply about what we can interpret from this analysis?

Answer key

A strong answer: *(i) The headline assumes that screen time before bed causes shorter sleep. (ii) The causation could run the other way. People who have trouble falling asleep may use their phones while they lie awake, or stay on screens because they do not expect to fall asleep for a while. Then poor sleep would cause more screen time before bed, not the reverse, and the plot would show the same downward pattern. (iii) The data record both variables at the same time, with each adult choosing their own screen time, so the data show an association but cannot show the direction of causation, or whether either variable has any effect on the other.*

What makes it strong: *(i) names the assumed direction; (ii) gives a concrete process and says that the process would produce the same plot; (iii) draws the conclusion about what can be interpreted from the analysis.*

Also right: *For (ii), a specific third variable that raises screen time and lowers sleep, such as stress, shift work or a new baby, described so that it produces the downward pattern; (iii) then says that the data cannot separate that variable's effect from the effect of screens. A reverse-causation process with a third variable added: no penalty for the addition.*

Weaker answers: *"Correlation isn't causation" without a specific process. "Screens do steal sleep, the plot shows it" takes the headline's direction for granted. A process that does not produce the downward pattern (for example, "some people just need less sleep") does not answer (ii).*

Rubric (4 points): *(i) 0.5 point: the assumed direction, screens cause less sleep. (ii) 2 points: a specific process, reverse causation or a named third variable (1.5), that would produce the downward pattern (0.5). (iii) 1.5 points: the data show an association but cannot show the direction of causation, or cannot separate the third variable's effect (1.5). "Correlation isn't causation" with no process: 0.5 in all. A process that does not produce the pattern: 0 for (ii).*

The idea to carry with you: *When two variables go together, consider both directions of influence, and any third variable that could move both variables. The variable treated as the outcome can be the cause of the "explanatory" variable.*

6.4 Testing the headline

Which study could show whether screen time before bed reduces sleep? Choose one, and give your reason.

A. Survey a random sample of 2,500 adults instead of 25 about their screen time before bed and their sleep.
B. Compare each adult's sleep on nights they chose to use screens before bed with their sleep on nights they chose to avoid screens.
C. Ask volunteers to choose whether to give up screens before bed for two weeks, and compare them with volunteers who keep their habits.
D. Randomly assign volunteers to avoid screens in the hour before bed or to feel free to use them, for two weeks; compare their sleep.

Answer key

Answer: *D.*

Why: *Random assignment, not the volunteers, decides who gives up screens, so the two groups are alike, apart from chance differences, in everything else that affects sleep, including how well each person slept before the study. A difference in sleep between the groups can then be attributed to the screens. Asking volunteers to give up screens for two weeks is also harmless.*

Why not the others:

- *A is a random* sample, *not random* assignment. *A random sample lets the estimate describe all adults, and a larger one makes it more precise, but the adults still choose their own screen time, so the association can still come from the effect of sleep on screen time. More data of the same kind cannot show the direction of causation.*
- *B compares each adult with themselves, which removes differences between adults, but the adults choose the nights: they might avoid screens on nights they already feel sleepy.*
- *C lets volunteers choose, so those who give up screens may be the ones who already sleep better or care more about sleep.*

Weaker answers: *D with "random is fair" alone has the right choice and half the reason. A full answer says that the assignment, not the person, decides who gives up screens, and names the flaw in B and C, the participants' own choices.*

Rubric (3 points): *D (1 point). Reason (2 points): random assignment, not the person, decides who gives up screens, so the groups are alike apart from chance (1.5); the flaw in each other option, A because a random sample is not random assignment, B and C because the choice is the participants' own (0.5). "Random is fair" alone: 0.5 of 2.*

The idea to carry with you: *To learn whether a treatment causes an outcome, let chance, not the participants, decide who receives the treatment.*

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

1. Before computing a summary, identify the cases in the table and the meaning of each variable's values (3.1).
2. Read a chart's axis before comparing its bars (3.2).
3. Separate "no effect shown" from "shown to have no effect" (5.4).
4. Expect larger samples to give less variable estimates, and know that sample size does not remove the bias from a poorly chosen list (4.1, 4.2).
5. Judge a result by its position among the results of chance alone (5.1, 5.2).
6. Expect false positives among many comparisons, and report the number of comparisons (5.3).
7. Expect a repeated study to give a somewhat different result, and compare results rather than verdicts (5.4).
8. Predict from the overall pattern (6.1, 6.2).
9. Know what can produce a pattern besides the headline's cause, and which designs allow a causal conclusion (6.3, 6.4).