

Chapter 15 Solutions

Making Sense of Social Data

Mark Paradis, Ph.D.

September 2, 2026

These are the worked solutions to the *end-of-chapter* exercises for Chapter 15 of *Making Sense of Social Data*. Solutions to the exercises inside each section are in the book itself, printed immediately after the exercises they answer.

Do the exercise before you read the solution. A solution read first is just more text to read; a solution read second is a check on your reasoning, and only the second one teaches you anything.

The complete set for every chapter is at <https://www.markparadispolitics.com/books/making-sense-of-social-data/online/solutions/>.

Chapter 15 Comparing Means

Solution to Exercise 15.E1. (a) A *paired samples t*-test. The same people are measured twice, before and after, so each participant supplies two linked scores and the two columns are not independent.

(b) One difference is one participant's after score minus their before score. With 45 participants there are 45 differences, so $n = 45$ and $df = 44$ — not 90.

(c) $H_0 : \mu_D = 0$, the mean change in confidence is zero. $H_1 : \mu_D \neq 0$, it is not. Two-tailed unless the direction was fixed before the data were collected; a programme can in principle reduce confidence as well as raise it, by showing participants how much they did not know.

(d) The test establishes whether the average confidence score changed by more than sampling error would explain over the twelve weeks. It does not establish that the programme caused the change. With no control group, anything else that happened over those twelve weeks is a rival explanation — the labour market improving, the season changing, participants simply becoming more familiar with the questionnaire. Chapter 3 sets out what a comparison group buys, and it is exactly this.

Solution to Exercise 15.E2. (a) The *Welch* version. The two standard deviations are 12.1 and 16.4, which is a substantial difference in spread, and the two groups are unequal in size. That is the combination the pooled estimate handles badly. In practice Welch is the right default whatever the variances look like (§15.3.3).

(b) The Welch standard error is

$$SE = \sqrt{\frac{12.1^2}{55} + \frac{16.4^2}{48}} = 2.87$$

and the statistic

$$t = \frac{74.3 - 78.9}{2.87} = -1.60$$

on 85.5 degrees of freedom, giving $p = .113$. At $\alpha = .05$, reject H_0 : the two classroom types differ in mean achievement.

(c) Cohen's $d = (74.3 - 78.9)/14.26 = -0.32$, a small-to-medium effect. The denominator is the *pooled* standard deviation, not the unpooled standard error the Welch test used: an effect size is a difference measured in the units people vary in. The 95 per cent interval for $\mu_1 - \mu_2$ runs from -10.32 to 1.12 points. Note that it excludes zero, which is the same verdict the test gave.

(d) It costs the causal claim entirely. Students were not assigned to classroom types, so the two groups may differ in prior attainment, in the schools that adopted the technology, in the teachers who volunteered to use it, and in anything that led

families to one classroom rather than the other. The test establishes a difference between two groups of students; it says nothing about what the technology does. Compare §15.3.4, where the same arithmetic supports a causal reading only because assignment was random.

Solution to Exercise 15.E3. (a) **Step 1.** $H_0 : \mu = 3.50$; $H_1 : \mu \neq 3.50$, two-tailed.

Step 2. A random sample is assumed; observations are independent; $n = 400$ is far above any threshold at which the central limit theorem is in doubt. σ is unknown, so the test is t .

Step 3. $\alpha = .05$, two-tailed, $df = 399$.

Step 4.

$$SE = \frac{0.91}{\sqrt{400}} = 0.045, \quad t = \frac{3.62 - 3.50}{0.045} = 2.64$$

Step 5. $p = .009$, which is below $.05$. Reject H_0 .

Step 6. Mean self-rated health in this sample differs significantly from the national average.

(b) $d = (3.62 - 3.50)/0.91 = 0.13$.

(c) It is detectable, and it is very small. The difference is 0.12 of a point on a five-point scale, and $d = 0.13$ is below the conventional threshold for a small effect. With $n = 400$ the study has enough precision to distinguish a difference this small from zero, which is what the significant p -value reports — and nothing more.

A headline would report the p -value, because “significant difference in health found” is a sentence and “the two groups differ by about a tenth of a point” is not. The second is the one that describes the finding. This is the gap between statistical and practical significance, and it widens as samples grow.

Solution to Exercise 15.E4. (a) Because a hypothesis test is a decision rule applied to a sample, and a sample is not the population. The rule is built to be wrong a controlled fraction of the time: at most α of the time when H_0 is true, and — the part that gets less attention — as often as the design’s power allows when H_0 is false. Nothing in the procedure guarantees a correct answer on any single occasion. In §15.2.6 the null is false every time, and the test still fails to reject in 32% of draws.

(b) Neither, necessarily. Two correctly conducted studies of the same false null will disagree whenever one draw lands in the rejection region and the other does not, which at this power is a common event rather than a strange one. Looking for the mistake is the wrong response; looking at both intervals, and at how large each study was, is the right one.

(c) Decide the sample size in advance, from the smallest effect that would matter and the power required to detect it — a power analysis. It cannot be fixed afterwards for two reasons. A study that has already produced a non-significant result cannot

distinguish “no effect” from “too small to see”, because that is precisely what it failed to measure. And adding cases until a result becomes significant, then stopping, changes the error rate of the test: the reported p -value no longer means what it says.

Solution to Exercise 15.E5. (a) The second. It reports $d = 0.45$ with an interval from -0.20 to 1.10 , so a reader knows the estimated size of the effect and the range the data are consistent with. The first reports only that a difference was detectable in a sample of 200, which is a fact about the study.

(b) No. The second study’s interval comfortably contains the first study’s likely effect size, and with $df = 38$ it is wide enough to contain almost anything modest. Two results are in conflict when their intervals do not overlap, not when one crosses a significance threshold and the other does not.

(c) Something of this form:

Wellbeing scores were higher in the treatment group ($M = \dots$, $SD = \dots$) than in the control group ($M = \dots$, $SD = \dots$), $t(198) = 2.14$, $p = .03$, $d = \dots$, 95% CI [$\dots, \dots$].

The two group means and standard deviations, the effect size, and the interval. Everything a reader needs to judge the size of the finding and to combine it with anyone else’s.

Solution to Exercise 15.E6. Answers will vary with the article chosen. A complete response works through all five questions of §15.5 against the article’s own text, and quotes the article where it makes a claim.

Three things are worth knowing before you start. Papers frequently describe a design in the methods section and report a test that does not match it, and the mismatch is usually visible from the degrees of freedom alone: a paired design on 30 participants gives $df = 29$, and $df = 58$ means the pairing was ignored. Effect sizes are reported far more often in psychology than in economics or political science, so the answer to (c) may say more about the discipline than about the authors. And an article reporting a single significant comparison rarely says how many were run — the honest answer to (d) is often “the article does not say”, which is itself a finding about the article.

A defensible answer to (e) can be either yes or no. What makes it defensible is that it names which of the four previous answers drove it.