

Chapter 3 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 3 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 3 Introduction to Research Design

Solution to Exercise 3.E1. (a) IV: eating breakfast — as implied, a binary (eats breakfast or does not), though frequency would be better. DV: grades, presumably a GPA or test score.

(b) H_1 : Students who eat breakfast have higher grades than students who do not. H_0 : There is no difference in grades between students who eat breakfast and students who do not.

(c) Two confounders, each with a mechanism.

Household income. Families with more resources are likelier to have breakfast available and a routine that makes time for it, and their children have more books, quieter study space, tutoring, and better-resourced schools. Income produces both the breakfast and the grades.

Household organisation or parental involvement. A household that reliably produces breakfast is likely to be one that enforces bedtimes, checks homework, and attends parent evenings. The breakfast is a marker of the household, not a treatment.

Sleep is also defensible: students who get up in time to eat are students who went to bed earlier, and sleep affects academic performance directly.

(d) A randomised controlled trial. Recruit students, and assign at random to receive a free breakfast programme or not. Randomisation makes the groups comparable on income, household organisation, sleep, and everything else measured or not. Compare grades at the end of the term.

There is a practical difficulty worth noting: you cannot stop the control group from eating breakfast at home, so the comparison is really between offered and not offered rather than eating and not eating. That is a real design — it is how school breakfast programmes are actually evaluated — but the question it answers is slightly different from the one the headline asked.

Solution to Exercise 3.E2. (a) *Spurious.* Z = national income or development. Wealthier countries have higher literacy and more newspaper readership, and also better maternal healthcare, sanitation, and nutrition, which reduce infant mortality. Nothing about reading a newspaper affects whether an infant survives.

(b) *Genuine causal, probably — but with reverse causation operating too.* Exercise plausibly improves mood through well-documented physiological routes. But happier people also have more energy and motivation to exercise, and depression reduces physical activity. Both arrows are real, which makes this reciprocal causation. Note that a cross-sectional correlation cannot apportion the two.

(c) *Reverse causation.* Cities hire more police *because* crime is high, not the other way round. Reading this as police causing crime inverts the sequence. This is the pattern to watch for whenever the independent variable is a policy response — responses are allocated to problems.

(d) *Confounded, and partly genuine.* Note-taking probably does help, through rehearsal and retrieval. But students who take more notes are also likelier to attend regularly, pay attention, and be motivated — and those affect exam performance independently. The honest answer is that some of the association is causal and some is not, and the raw correlation overstates the effect of note-taking.

The point across all four: “spurious” and “causal” are not always a clean binary. Real relationships are frequently partly both, which is what makes them hard.

Solution to Exercise 3.E3. (a) *Research question:* Does attending university reduce prejudice toward out-groups? H_1 : Adults who have attended university report lower levels of prejudice toward out-groups than adults who have not. H_0 : There is no difference in prejudice between adults who have and have not attended university.

(b) IV: university attendance — binary, or better, years of post-secondary education completed. DV: prejudice, which needs care. A validated scale such as a social distance measure, or feeling thermometers toward named groups. Self-report on this topic is vulnerable to social desirability (Chapter 2), which is worth flagging when you propose the measure.

(c) The ideal design is an experiment, and it cannot be run — you cannot randomly assign people to attend university. The best feasible design is a longitudinal panel measuring prejudice before entry and again after graduation, with a comparison group of similar people who did not attend.

(d) Two confounders. *Family background*: educated, wealthier families are both likelier to send children to university and likelier to hold and transmit less prejudiced attitudes. *Prior attitudes*: young people already lower in prejudice may be likelier to choose university, or to choose particular programmes — selection rather than effect.

(e) A longitudinal design measures prejudice *before* university, which addresses the second confounder directly: if you know each person’s starting level, you can ask whether it changed rather than whether it differs. It also establishes temporal precedence.

What it does not fix: family background still differs between those who attend and those who do not, and anything that affects both attendance and attitude change over the same period remains confounded. Longitudinal data removes stable differences between people; it does not remove differences in their trajectories.

Solution to Exercise 3.E4. (a) Cross-sectional — 2,400 adults measured once.

(b) *Covariation*: satisfied. $r = .31$ with $p < .001$ establishes an association. Note that $r = .31$ is moderate; cohesion accounts for a modest share of variation in health.

Temporal precedence: not satisfied, and cannot be by this design. Both were measured at one moment. The reverse is plausible — people in better health get out

more, know their neighbours, and perceive their neighbourhood as more cohesive.

Elimination of alternatives: not satisfied. No confounders addressed.

(c) No. The conclusion is a causal claim — *building* cohesion would *improve* health — and the design supports only an association. The policy implication is what makes the overreach consequential: if the relationship is confounded, a programme that increases cohesion will not improve health, and the money is spent for nothing.

(d) Two confounders. *Neighbourhood income or deprivation:* affluent areas have more stable residents, more community amenities, and better health through diet, healthcare access, and working conditions. *Individual health status:* sick people participate less in community life and report worse health, producing the association through the DV itself.

Residential stability, age, and home ownership are all defensible alternatives.

(e) A longitudinal panel following the same adults over several years would establish sequence and allow each person to serve as their own comparison. Stronger still would be a natural experiment — some external change in cohesion unrelated to health, such as a neighbourhood redevelopment — though these are rare and rarely clean.

Solution to Exercise 3.E5. (a) *Data needed.* Education spending, ideally as a share of GDP rather than absolute dollars, so countries of different size are comparable (§2.6.5). GDP per capita. Both for the same set of countries, and — crucially — measured at more than one point in time if the claim is to be tested rather than described. Sources: World Bank or OECD.

(b) *Design.* A cross-national comparative design, and if measured at one point, cross-sectional. This supports description and correlation, and not causation.

(c) *Likely causal?* Partly, at best, and the alternatives are strong.

Reverse causation is the most serious. Rich countries can afford to spend more on education. Since spending is usually measured as a share of GDP, a wealthier country spending the same share spends far more per student — and the direction runs from wealth to spending as readily as the reverse.

Confounding. State capacity, political stability, and rule of law all plausibly raise both public spending and economic output. So does an existing skilled population.

Timing. Education spending affects the economy through the productivity of people educated by it, which is a lag of fifteen to twenty-five years. A contemporaneous correlation is measuring the wrong interval entirely.

(d) *Moving to causation.* Establish sequence with panel data over decades, so that spending changes can be shown to precede growth changes. Control for the measurable confounders. Look for a natural experiment — an education reform driven by something unrelated to the economy. And specify the mechanism, since “spending” covers teacher salaries, buildings, and university subsidies, which need

not work the same way.

This is among the most-studied questions in development economics and it remains contested, which is itself informative about how hard causal inference is from comparative data.

Solution to Exercise 3.E6. Your question will be your own, so this solution sets out what each of the six elements has to do, and then gives one worked specimen you can check yours against.

1. *The unit of analysis, stated.* This is the commonest omission. Individuals, households, neighbourhoods, and countries are different studies even when they use the same two variables.

2. *Operationalisation that is specific.* “Measure wellbeing” is not an operationalisation. Naming a scale, a set of items, or an administrative indicator is.

3. *A directional H_1 , and an H_0 that is the absence of a relationship* rather than its reversal. H_0 says there is no difference; it does not say the difference runs the other way.

4. *A justification that names a mechanism.* This is the element most often left thin. “Prior research suggests a link” is empirical justification only. An account of *how* the cause would produce the effect is what makes a hypothesis worth testing, and it usually generates further predictions you can check.

5. *A design matched to the question, with one threat acknowledged.* Every design in §3.2 leaves something open. If you cannot name a weakness in yours, you have not yet found it rather than avoided it.

6. *A confounder that is plausible and justified.* It has to affect both the independent and the dependent variable, and you have to say why it would.

Keep the whole thing feasible. A proposal requiring a twenty-year panel of forty countries is not a design; it is a wish.

A worked specimen.

(1) Among CEGEP students in Quebec, does working for pay during the semester lower academic performance? The unit of analysis is the individual student.

(2) IV: paid hours worked per week during the semester, self-reported as a count. DV: term grade average, taken from institutional records rather than self-report, since students misreport grades in a direction that flatters them.

(3) H_1 : students who work more paid hours per week have lower term grade averages. H_0 : there is no relationship between paid hours worked per week and term grade average.

(4) Mechanism: time is finite. Hours worked are hours not spent on coursework, sleep, or attendance, each of which affects performance. The mechanism also predicts something further — the effect should be larger at high hour counts than at

low ones, since the first few hours come out of leisure and the later ones out of study.

(5) Design: a cross-sectional survey linked to grade records, which is feasible for a student project. Threat to internal validity: selection. Students who choose to work long hours may differ from those who do not in ways that affect grades on their own, so the design cannot separate the working from the kind of student who works.

(6) Confounder: household income. Students from lower-income households both work more hours and face other disadvantages — longer commutes, less quiet study space, fewer paid supports — that lower grades independently of the job.

Solution to Exercise 3.E7. (a) A *comparative* design — countries as units, compared with one another. Because it has three waves it is also a panel, though of countries rather than individuals.

(b) *Covariation*: yes, this the dataset can establish. A strong positive association between schooling and life expectancy is measurable.

Temporal precedence: partly. Three waves let you ask whether countries whose schooling rose earlier saw life expectancy rise later. That is more than a cross-section offers and less than a long panel.

Elimination of alternatives: no. The dataset contains a handful of variables, and any number of unmeasured national characteristics could produce the association.

So the honest claim is that education and life expectancy are associated across countries, and the association persists across waves — not that education raises life expectancy.

(c) National income is the obvious confounder: richer countries have both more schooling and longer lives. Checking Appendix B, *worldsim* *does* contain `gdppc`, so this one is measurable. Many others — health system quality, conflict, sanitation — are not, which is the point of the exercise: the confounders you can address are the ones somebody thought to measure.

(d) Three waves add the ability to look at change over time and to establish rough sequence. They do not provide randomisation, do not capture unmeasured confounders, and cannot rule out that some third national characteristic changed alongside both. Three observations is also a short series for detecting lagged effects.

(e) The figures will not match, and they are not supposed to. *worldsim* gives Canada a life expectancy above eighty-four years in the first wave and a little over eighty in the second — a fall of several years across a single wave, which no real country experienced over the period the World Bank covers. The schooling figures are in a plausible range but do not track the published series either.

The lesson is the one to carry through the rest of the book. *worldsim* was built so that education and life expectancy are strongly associated; the association is in the data because it was put there. A conclusion drawn from it is a conclusion about the simulation, and it tells you whether you have applied the method correctly —

which is what the dataset is for. It is not evidence about Canada, and no number in it should ever be quoted as though it were.

Solution to Exercise 3.E8. The article is yours, so what follows is what each part is asking for, and the pattern you are most likely to find.

(a) The gap to look for is between the headline’s verb and the study’s. “Linked to” in the body of the piece becoming “causes” in the headline is the pattern described in §3.5.4, and it is very common. Restating the claim precisely means naming the population, both variables as they were actually measured, and the strength of the relationship if it is reported.

(b) Many articles do not report the design at all. Saying so is a correct answer, not a failure to find one — and the absence is itself informative, because a claim whose design is undisclosed cannot be evaluated by a reader.

(c) Apply the four questions to this article rather than in general. “There could be a confounder” does not advance the analysis; naming a plausible one, and the mechanism by which it would affect both variables, does.

(d) The sentence to aim for has this shape: *this study shows that X and Y are associated among [population], and does not establish that X causes Y*. If you can write that sentence about your article, you have the skill this chapter was for.

One thing you will often notice: the article’s own text is more careful than its headline, and the researchers quoted inside it are more careful still. That chain of escalation — from researcher, to press release, to headline — is the ethics point from §3.5.