

Chapter 1 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 1 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 1 Introduction to Quantitative Methods

Solution to Exercise 1.E1. (a) The word doing the damage is *proves*. A study showing that teenagers who use social media more report more depressive symptoms has established an association. Three other explanations remain live: the causal arrow may run the other way, with depressed teenagers withdrawing into their phones; some third factor — sleep disruption, social isolation, family circumstances — may drive both; or the association may be an artefact of who ended up in the sample. Chapter 3 treats this properly.

(b) Supporting a causal claim requires design, not more data. A longitudinal study measuring both variables in the same people over years could at least establish which came first. An experiment randomly assigning some participants to reduced use would be stronger still, since randomisation makes the groups comparable on everything, measured or not. Both need validated instruments rather than a single self-report item, and both need a sample that represents some defined population.

(c) Several ethical issues are available. Research on minors requires parental consent alongside the young person's own assent. Asking about depressive symptoms may cause distress and creates a duty to provide support information. An experiment restricting social media access intervenes in participants' social lives, which needs justification proportionate to the knowledge gained.

Notice that (b) and (c) pull against each other: the design that would best support the causal claim is also the one raising the most serious ethical questions. That tension is real and recurs throughout the book.

Solution to Exercise 1.E2. *Quantitative design.* Question: does frequency of political content consumption on social media predict voting and other participation among 18–29 year olds? Data: a survey of a representative sample of that age group, measuring platform use, exposure to political content, and several forms of participation, with demographic controls. Strength: results generalise to the population sampled, and competing explanations such as education can be accounted for. Limitation: it captures what people report rather than what they do, and even a strong association will not establish direction.

Qualitative design. Question: how do young adults understand the relationship between their online lives and their political action? Data: twenty to thirty in-depth interviews, purposively sampled for variation in participation. Strength: it can uncover mechanisms and meanings a survey cannot anticipate — including reasons nobody thought to put on a questionnaire. Limitation: it cannot establish how common any pattern is.

These are not competing answers to the same question. The survey establishes that a pattern exists and how widespread it is; the interviews explain what is happening inside it. A mixed design running interviews first would improve the

survey, because it would supply the participants' own vocabulary for the questions.

Solution to Exercise 1.E3. Open-ended. Four criteria to check your answer against.

Are the three questions genuinely from different disciplines? Three variations on voting behaviour are one discipline three times. Aim for questions a sociologist, an economist, and a psychologist would each recognise as belonging to their field.

Are the questions answerable with this dataset? A question about change over time needs a dataset with multiple waves. A question about causal effect cannot be settled by cross-sectional survey data; say so rather than overclaiming.

Are the variables named specifically? This is where browsing the codebook pays off: name actual variables, or quote precise question wordings.

Is the analysis plan sketched sensibly? Knowing the technique's name is not required. "Compare average trust scores between people with and without post-secondary education" is a complete and correct answer at this stage.

Solution to Exercise 1.E4. (a) Three issues, at minimum.

Consent. Posts written for followers were not offered to researchers. Public visibility describes who *can* see something, not who it was intended for, and terms of service are not informed consent.

Sensitivity and harm. Mental health information carries stigma and can affect employment, insurance, and relationships. People posting during a disaster are, by definition, in distress and less able to consider how their words might travel.

Identifiability. A verbatim quotation is usually traceable through a search engine, and a handful of attributes — age band, city, occupation — often narrows a population to one person. "No names were used" is not anonymisation.

(b) Mitigations: report only aggregate patterns; paraphrase rather than quote; seek Research Ethics Board review even though the data are public; exclude accounts belonging to minors; and consider whether the research question could be answered with data whose subjects consented.

(c) Either answer is defensible; the reasoning is what matters. A defensible *yes* weighs the public benefit of understanding disaster mental health against low individual risk when only aggregates are reported, and accepts the mitigations above as conditions. A defensible *no* holds that consent cannot be dispensed with merely because obtaining it is inconvenient, and that vulnerability during a disaster strengthens rather than weakens the obligation. An answer that reaches a verdict without weighing the other side leaves the exercise half-done.

Solution to Exercise 1.E5. The argument contains a fair characterisation and a false conclusion, and the response should separate them.

What is fair. Post-positivism does accept that measurement is imperfect and that researchers' values shape what they study. That is an accurate description, not a concession extracted under pressure.

Where it fails. The conclusion assumes that without certainty there is no basis for comparing studies. That does not follow. Studies can be compared on grounds that have nothing to do with certainty: whether the sample represents the population it claims to, whether the measure has been validated, whether the analysis was specified before the data were seen, whether the finding replicates, whether the reported uncertainty matches the evidence. A study with a representative sample and a validated instrument is better than one with neither, and no appeal to certainty is required to say so.

The deeper point. Post-positivism replaces individual certainty with collective procedure. Objectivity is not a state a researcher achieves by being unbiased; it is what method, transparency, replication, and criticism by people with different priors produce between them. The argument treats objectivity as a property of individuals, which is exactly the assumption post-positivism abandons.

Agreeing with the original claim remains defensible, provided the response engages this counter-argument rather than restating the claim.

Solution to Exercise 1.E6. Defensible answers vary; the limitation column is the one that repays thought.

Income bracket (7 categories). Economics, sociology. Limitation: brackets discard information within each range, and the top bracket is open-ended, so the upper tail — where inequality lives — is invisible.

Trust in government (0–10). Political science. Limitation: “government” is ambiguous — the current administration, the civil service, or the system itself — and respondents may answer about different objects.

Country of birth. Sociology, geography, demography. Limitation: it records where someone was born and nothing about when they arrived, their status, or their attachment; it is often used as a proxy for things it does not measure.

Number of children. Demography, economics, sociology. Limitation: ambiguous as to whether it counts children ever born, currently living, or currently in the household — three different numbers.

Self-reported health (5-point). Public health, psychology, sociology. Limitation: it measures perceived health, and the same objective condition is reported differently across cultures, ages, and expectations.

Party identification. Political science. Limitation: the categories are country-specific and unstable over time, which makes cross-national and long-run comparison difficult.

The pattern worth drawing out: every one of these limitations is about what the measurement leaves out or blurs, not about the arithmetic applied afterwards.

Solution to Exercise 1.E7. Your answer depends on the countries you chose. The reasoning below applies whichever three you picked.

(a) The gaps will be substantial — several years in most cases — and this is by design rather than by error.

(b) The ordering should hold. A country with higher real life expectancy should have higher simulated life expectancy.

(c) Yes, and this is the point of the exercise. The book's data reproduces the *structure* of the world — which countries rank above which, and roughly how relationships bend — while deliberately misstating the *levels*.

(d) Three reasons; two of them are enough. First, simulated data has known ground truth: because the data was built rather than observed, the true population values are available, so an estimate can be checked against the right answer in a way no real dataset permits. Second, it is stable — real figures are revised, which would make exercises stop matching their answers. Third, licensing: reproducing real datasets inside a commercially sold book raises costs and restrictions that synthetic data avoids.

There is a fourth worth adding. Having verified the distortion yourself, you now have a concrete reason to distrust any number in this book as a fact about the world — which is exactly the habit the Preface asks for.

Solution to Exercise 1.E8. Open-ended. The point is to read a methods section closely, which is a habit worth building early.

(a) *What was measured, and how.* Identify the actual instrument — a specific scale, an administrative record, a set of survey items — rather than the concept. “They measured wellbeing” restates the variable; “they used a five-item scale asking about the past two weeks” describes the measurement.

(b) *Who was studied, and how selected.* Distinguish the population from the sample, and note the recruitment method. Studies are often silent on this; noticing the silence is part of the answer.

(c) *One thing it cannot tell you.* The most useful limits are the ones arising from the design rather than from the topic: that a cross-sectional study cannot establish sequence; that a sample drawn from one city cannot speak to a country; that a self-report measure cannot distinguish a real change from a change in willingness to report.

A methods section is often harder to read than the results, and often contains less detail than expected. Both are worth noticing.

Solution to Exercise 1.E9. 1. *Has trust in national governments declined across wealthy democracies over three decades?* (a) Quantitative. It asks about extent and change over time across many cases — exactly what surveys with repeated waves are for. (b) The unit of analysis is the country-year, though the underlying data are collected from individuals. (c) The obstacle is comparability: the question must have meant the same thing in every country and every decade, and translation,

changing political vocabulary, and shifting survey modes all threaten that.

2. *Why do some people who distrust their government nonetheless continue to vote?* (a) Qualitative, or mixed with a qualitative core. The question asks about reasoning and meaning, and the researcher cannot anticipate the range of answers well enough to write closed questions. (b) The unit is the individual. (c) The obstacle is that people are not always reliable narrators of their own motives, and post-hoc justification is hard to distinguish from actual reasoning.

3. *Does a public information campaign increase trust in a local health authority?* (a) Quantitative, and experimentally if possible — “does X increase Y” is a causal question, and randomisation is the strongest available answer. (b) The unit is the individual, or possibly the neighbourhood if the campaign is delivered geographically. (c) The obstacle is contamination: people in the control group may see the campaign anyway, and trust may be moving for reasons unconnected to it.

The pattern worth drawing out is that the form of the question, not the topic, determines the approach. All three are about trust in government; they need three different designs.