A Field Guide to Missing Data

What economists do when the data they need aren't there

Beatriz Gietner
DiD Digest, diddigest.xyz

August 2026

Introduction

Empirical work often starts with an object we can't observe. Sometimes that object is a variable or an observation. Sometimes it is a counterfactual, a comparable definition, a price, a target population, an assignment mechanism, a joint distribution, or a reproducible computational record. I use *missing data* in this broad sense[1]. There's a reason I think we should stop and think a bit deeper about these "absences": they require different responses. An economist might report a set instead of a point, stress an estimate under a departure from an assumption, reconstruct an unavailable object, combine information across sources, generate a reference distribution, change the target question, regularise a sparse model, or improve the reporting and audit trail. These are different inferential activities, and because they are different, in nature and in substance, they don't become interchangeable when a results table places them under one "robustness" heading.

So to get more out of the imperfect situation we find ourselves in, the first question we should ask is: what exactly is unavailable? This requires a lot of training, actually, and the next ones aren't easier either. After we've thought about what is unavailable, one way "out" is to ask: what side information is available? Which assumption links it to the missing object? What can the proposed exercise establish under that assumption? What remains unknown afterwards? And has the target estimand (the quantity we're trying to estimate) changed along the way?

With all of this in mind, I tried my best to come up with a field guide organised around 22 recurring problems. Several entries below contain exercises that do different jobs, and I keep them separate whenever the choice between them changes what the analysis can claim: a bound, a reconstruction and a stress test answer to different things, even when they address the same problem.

Here's a list of the entries covered in this document:

- P01. A relevant confounder is unobserved
- P02. The variable you want is represented by a proxy
- P03. Some ordinary observations or items are missing
- P04. Outcomes disappear through attrition or selection
- P05. A continuous variable is measured with error

[1] I'm using missing data more broadly than usual, although the basic idea has precedents. Howe et al. (2015) treat confounding, selection and measurement error as missing-data problems; Edwards et al. (2015) describe measurement error as missing information within the potential-outcomes framework; and Ding and Li (2018) develop causal inference as a missing-potential-outcomes problem.

- P06. Either treatment or outcome status is misclassified
- P07. A price or quality adjustment is unavailable
- P08. The counterfactual is unavailable
- P09. Inputs or computational objects are inaccessible
- P10. Only aggregates are observed
- P11. Definitions or classifications change
- P12. The sample is small or the event is rare
- P13. There is little overlap or little identifying variation
- P14. Records describing the same units can't be linked exactly
- P15. Variables sit in different datasets
- P16. Values are censored, top-coded or altered for disclosure
- P17. The official outcome arrives too late
- P18. The study population differs from the target
- P19. Treatment timing is uncertain or anticipated
- P20. One unit's treatment affects another unit
- P21. The available measure doesn't settle the concept
- P22. Data are revised after first release

P01. A relevant confounder is unobserved

We shall make this easier to understand with a familiar setup: we want to know whether a job-training programme raises earnings, and people chose whether to enrol. One worry for us would be "motivation": one would assume more motivated people are more likely to sign up and more likely to earn more anyway, and no .csv has a column called "motivation". That's the missing object here: the part of who-gets-treated that's related to the outcome but that our controls don't capture. Confounding has other responses, and a convincing instrument, a discontinuity or randomised variation can change the design itself (Angrist and Pischke, 2010), but this entry is about what we can learn when the original observational comparison is kept in place.

Since we can't observe motivation directly (and don't have a "suitable" proxy), the best we can do is ask the "what if" question: how strong would the link between motivation, enrolment and earnings have to be before our estimate stopped meaning what we say it means? That's what sensitivity calculations do. Robustness values and coefficient-stability exercises (Cinelli and Hazlett, 2020; Oster, 2019) put numbers on that question. Oster's version asks it in a more specific way: suppose the unobservables (e.g. the "motivation") are related to enrolment in the same way as the observables we do observe in the data, like education and work history. If controlling for the things we can see barely moves the estimate while those controls explain a good share of the outcome, it would take an implausibly strong hidden factor to overturn it; if the estimate jumps around as we add controls, a hidden factor could plausibly finish the job. Both halves of that condition are needed since stable coefficients alone can be misleading: controls that carry little information also leave the estimate unmoved, without telling us anything about

the confounder (Pei et al., 2019). To run this we have to declare two inputs: δ[2], which says how strong the hidden selection is allowed to be relative to the observed selection, and R-max[3], which caps how much of the earnings variation a complete model could ever explain. The calculation won't choose those values for us, and what it returns is a threshold that stresses the estimate, not an estimate of the missing confounding itself.

Rosenbaum sensitivity analysis asks a similar question from the assignment side, and it's an easy one to over-read. Take two people who look identical on every variable we observe. Its Γ parameter[4] says how much more likely one of them could be to enrol because of the hidden factor: Γ of 2 means one could face twice the odds of treatment. The procedure then reports whether our conclusion would survive that much hidden imbalance. So what it bounds is the inference under that specified hidden-bias model rather than a causal parameter in general (Imbens, 2003).

A negative control borrows a habit from epidemiology: find an outcome the programme couldn't possibly affect, and check whether we "find" an effect there anyway. Earnings in the years before the programme existed are a natural candidate. If our design shows the training "raising" earnings that were paid out before anyone was trained, we haven't discovered time travel; we've diagnosed a bias pattern, most likely the motivated people pulling ahead in any period (Lipsitch et al., 2010). The catch is that the diagnosis is only as good as our claim that the control outcome really is untouchable by the treatment. And going further, actually using negative controls to identify the causal effect, is a different exercise with its own requirements: proxy conditions,

[2] δ is the coefficient of proportionality: a number that says how strong selection on unobservables is allowed to be relative to selection on observables. If $\delta = 1$, the factors we can't see (like motivation) are related to programme take-up just as strongly as the factors we included in the regression (like education and work history); $\delta = 2$ would mean twice as strongly, and $\delta = 0$ takes us back to assuming no hidden selection at all. The idea comes from Altonji et al. (2005), who were studying Catholic schools and argued that if the variables we observe are effectively a random subset of all the factors that drive selection, then how much the observables shift our estimate is informative about how much the unobservables could. Oster (2019) turned this into the calculation applied papers now report, where the usual output is the δ that would drive the estimate to zero: if it would take hidden selection stronger than everything we measured combined ($\delta > 1$), many authors read the result as safer, though that reading still depends on how good our observables were in the first place. A search tip: the useful terms are "proportional selection", "coefficient stability", and the command names, `psacalc` in Stata and `robomit` in R.

[3] R-max is the R-squared of a hypothetical regression we can never run: outcome on treatment plus every relevant control, the observed ones and the unobserved ones together. It answers the question "how much of this outcome could we ever explain if we had all the right variables?", and the sensitivity calculation needs it because the room left between our actual R-squared and R-max is exactly the room where a hidden confounder could live. It can't be lower than the R-squared we actually got, and for most outcomes it shouldn't be 1 either, since part of the variation in something like earnings is noise, luck and measurement error that no variable will ever explain. Oster (2019) suggests a default of 1.3 times the observed R-squared, calibrated so that the method would validate the majority of a sample of results from randomised experiments, but that default is a starting point for an argument, and the right move is to justify a value for your own outcome (or show a range) rather than copy the 1.3. When searching, write it out as "Rmax" together with "Oster", or look for "maximum R-squared" and the "1.3 rule" for the discussions around the default.

[4] Γ is Rosenbaum's sensitivity parameter, and the easiest way to read it is through a matched pair. Take two people who look identical on every covariate we observe. If assignment were as good as random given those covariates, both would have the same odds of ending up treated, and Γ measures how far from that ideal we're willing to imagine: $\Gamma = 1$ means no hidden bias at all, while $\Gamma = 2$ allows the hidden factor to double the odds of enrolment for one of them relative to the other (odds, and not probability, is the technically right word here). The procedure then asks how our p-value would look in the worst case allowed by each value of Γ, raising Γ until the worst-case p-value crosses our significance level, and that value of Γ is what gets reported: a study that is "insensitive up to $\Gamma = 1.8$" keeps its conclusion as long as no hidden factor comes close to doubling the odds of treatment. The method comes from Rosenbaum (1987) and is developed in his book *Observational Studies* (2002, second edition), whose "Sensitivity to Hidden Bias" chapter is the standard introduction. Two reading habits are worth keeping: the bound applies to the inference (the p-value or CI) under this specific model of hidden bias, and a large Γ doesn't certify the design, it just says the rejection survives that much imagined imbalance. When searching, the terms that work are "Rosenbaum bounds", and the implementations are `rbounds` in R and Stata plus `sensitivitymv` and `sensitivitymw` in R for matched sets.

conditional independence, and rank or completeness conditions (Miao et al., 2018).

This family also has two more members worth talking about. When economic knowledge supports a sign or relative-magnitude restriction on the omitted bias, the output can be an identified set rather than a threshold (Chalak, 2019; Krauth, 2016). And for ratio measures, epidemiology has what they call the E-value: it asks how strongly an unmeasured confounder would have to be associated with both treatment and outcome to explain away the observed association, which is one more threshold, and not evidence that no such confounder exists (VanderWeele and Ding, 2017).

Whichever of these we run, none of them observes motivation for us, and none of them proves exchangeability, meaning that enrolled and non-enrolled people are comparable once we condition on the controls. They tell us how fragile our conclusion is to the confounder we're worried about, which is valuable, but different from removing the worry.

P02. The variable you want is represented by a proxy

This time the missing object is the variable itself. Staying with our training study: suppose earnings are self-reported, and what we actually want is what employers paid. People round, forget, and sometimes flatter themselves, so the column called "earnings" in our dataset is really a proxy for earnings. The same situation shows up everywhere once we look for it: test scores standing in for learning, self-reported health standing in for health, patents standing in for innovation, etc.

The standard way out is a validation sample: a dataset where we observe both the proxy and a better benchmark for the same people - like a subset of our survey linked to tax records. A replicate measure or a reliability study can recover parts of the noise, while a linked benchmark can also reveal systematic bias, so these tools give us different kinds of side information rather than substitutes for one another[5]. With one of these in hand, we can estimate how the proxy relates to the target, which is our model of the error process[6], and then use that estimated relationship to correct the analysis in the main sample.

This move rests on an assumption we should say out loud: it needs both the measurement process and the relationship we estimated to transport from the validation setting to ours, meaning they hold in our sample too (Bound et al., 2001). If people misreport differently in

[5] A validation sample, a linked benchmark, a replicate measure and a reliability study sound interchangeable, but they observe different things, and it's worth keeping them apart. A *validation sample* observes the proxy and a better benchmark for the same people, so it can estimate both the noise and any systematic gap between them. A *linked benchmark* is the administrative version of the same idea, like survey answers linked to tax or payroll records. A *replicate measure* is the same instrument applied twice, asking the same person their earnings in two interviews: comparing the answers tells us how noisy the instrument is, but if both answers share the same bias (everyone rounding their salary up), the replicate can't see it, because the bias cancels in the comparison. A *reliability study* is what statistical agencies run when they re-interview a subsample and publish the resulting reliability ratio - which is the share of the measure's variance that is signal rather than noise. The practical rule: replicates and reliability studies can tell us about noise, while only a benchmark can tell us about bias.

[6] The textbook error model is *classical measurement error*: the gap between the proxy and the truth is random noise, unrelated to the true value and to everything else in the analysis. It's the assumption behind the familiar claim that mismeasurement pulls estimates toward zero (more on when that claim holds in P05). The reason to be suspicious of it is that when economists actually checked, real survey errors turned out not to be classical. Bound and Krueger (1991) compared what people told the Current Population Survey with their Social Security payroll records and found errors that were *mean-reverting*: high earners tend to underreport and low earners to overreport, so the error is correlated with the truth, which is exactly what the classical model rules out. Bound et al. (1994) found the same pattern in the PSID Validation Study, a survey run on workers of a firm that shared its payroll records. This is why the transport assumption in the main text is doing real work: a correction built on the wrong error model can move an estimate in the wrong direction, not just by the wrong amount. When searching, the useful terms are "errors-in-variables", "classical measurement error", "mean-reverting measurement error", "reliability ratio" and "PSID Validation Study", and the survey covering all of it Bound et al. (2001).

the validation study than in our survey (because the years, the interview and/or the people themselves differ), then the correction we imported was calibrated for a different problem, and the apparent fix has just moved the problem from one dataset to the other. This is not ideal. One practical detail follows: for a regression correction, the validation records need to hold the proxy, the benchmark, the outcome and the relevant covariates together, and if the benchmark itself is imperfect, its error has to be modelled too, with the uncertainty from the calibration step carried into the final interval (Bound et al., 2001).

We should also be clear about what the correction returns when it does work: a regression parameter under the assumed error model. We can recover, let's say, the average effect of training on true earnings, and still have no idea what any particular person truly earned. And none of this tells us whether the proxy measures the concept we care about in the first place: a validation sample can tell us how noisy our earnings measure is, but it can't tell us whether earnings were the right thing to measure.

P03. Some ordinary observations or items are missing

Back to our training survey: some people answered everything except the earnings question, and a few interviews never happened at all. The missing object here is ordinary, a value that should have been sitting in our analysis sample, and the response to it is where most of us first met the phrase "missing data".

There are three standard routes here: imputation, weighting and observed-data likelihood, and the first two work in opposite directions. Multiple imputation fills each "empty space", several times, with plausible values drawn from a model of the data[7]. Inverse-probability weighting doesn't fill anything: it makes the people we did observe count for more so each respondent also stands in for the similar people who didn't respond. The third route, observed-data maximum likelihood, neither fills nor reweights but writes down a likelihood for the data we saw and integrates over the values we didn't, and it inherits the same ignorability and modelling assumptions as the other two (Little and Rubin, 2019). (*The simplest response of all - dropping the incomplete cases - is transparent but it pays in precision and can easily change the target to the kind of person who has complete records.*) All three routes can support valid inference, but only under an assumption called missing at random, or MAR[8], and the fine print on MAR is where the argument lives: it's always relative to the information we observe. Given what we see about a person, their age, education and work history, the chance that their earnings answer is missing can't depend on the earnings themselves. Whether that holds can't be checked from the observed data alone (Rubin, 1976; Robins et al., 1994); adding more variables to the model can make it more plausible, but never verified.

[7] Why fill each space several times instead of only once? Because filling once with our best guess pretends we knew the value all along, and the analysis then reports standard errors as if there had been no space, which overstates our certainty. Rubin's solution was to draw several plausible values from a model that reflects how uncertain we are, produce several completed datasets, run the analysis on each, and combine the results with a set of formulas known as Rubin's rules, so the disagreement across the completed datasets flows into the final standard errors. The foundations are in Rubin (1976) and his 1987 book *Multiple Imputation for Nonresponse in Surveys*. The implementations most people use are `mice` in R and `mi impute` in Stata, and those are also the best search terms together with "multiple imputation" and "Rubin's rules".

[8] The taxonomy comes from Rubin (1976), and the names are famously confusing, so here they are in plain words. MCAR, or missing completely at random: the spaces are unrelated to anything, as when a page of questionnaires is lost in the office. MAR, or missing at random: given the variables we observe, the chance of a space doesn't depend on the missing value itself, so younger respondents may skip the earnings question more often, but among people with the same observed characteristics, skipping is unrelated to earnings. MNAR, or missing not at random: the space depends on the value hiding in it even after we condition on everything observed, as when high earners "prefer not to say". The names mislead because MAR doesn't mean "missing randomly"; it means "random enough once we condition on what we see". Search terms that work: "missing at random", "ignorability", and for the MNAR case "nonignorable nonresponse".

There's also a "belt-and-braces" version, augmented inverse-probability weighting, which combines a model of the outcome with the model of who responds and stays consistent when either of the two is correctly specified (Bang and Robins, 2005). This property, double robustness[9], is a genuine insurance against picking the wrong model, but the insurance operates inside the maintained MAR assumption: it gives us two routes through model specification, and it doesn't repair data that are missing not at random (when the chance of a space depends on the value hiding in it), nor a positivity problem (when some kinds of people never respond at all, so nobody is available to stand in for them).

Imputation carries one more requirement that sounds bureaucratic and turns out to be important: the imputation model has to be compatible with the analysis we are planning to run[10]. If our analysis studies an interaction (e.g., training effects that differ by gender, and the imputation model doesn't include it), we've quietly built the absence of that interaction into the completed data before the analysis even starts. With chained equations - which is the popular variable-by-variable default - keeping the two compatible may need a substantive-model-compatible imputation scheme rather than the out-of-the-box settings (Bartlett et al., 2015).

Diagnostics have limits here too. We can spot completed values that look implausible or weights that "explode", which diagnoses instability or thin support, though it doesn't tell us which model or assumption caused it, and no diagnostic computed on the observed data verifies the missingness assumption itself. The assumption stays an argument, which is why the choice of variables in the imputation or response model is a substantive decision and belongs in the paper's text rather than in a replication file.

P04. Outcomes disappear through attrition or selection

Our training study has one more trick to play on us. At the follow-up interview, some people have left the study, and wages exist only for people who are working. This looks like P03, but it's a different thing because now the treatment itself is "editing" the sample: training changes who finds a job, and having a job determines whose wage we get to see. If we naively compare the wages of trained workers with the wages of untrained workers, we're mixing two things: the effect of training on wages AND the effect of training on who shows up in the wage column at all.

There are four responses here, and they answer different questions. The first borrows from P03: if dropping out is ignorable given everything we observed along the way, inverse-probability-of-censoring weights let the people who stayed stand in for the people who left, with the same MAR and positivity requirements as before, and no help against dropout that depends on the missing wages themselves (Robins et al., 1994). The second is to give up the point estimate and

[9]The definition is worth quoting, because the original authors say it more cleanly than most textbooks I've checked: "an estimator is doubly robust (DR) or doubly protected if it remains consistent when either a model for the missingness mechanism or a model for the distribution of the complete data is correctly specified. Because of the frequency and near inevitability of model misspecification, double robustness is a highly desirable property" (Bang and Robins, 2005). The augmented weighting estimator itself goes back to Robins et al. (1994). Positivity, the other requirement in the main text, is the condition that every kind of person in the analysis has some chance of being observed; weighting can stretch respondents to cover similar non-respondents, but it can't conjure information about groups with no respondents at all. Search terms: "AIPW", "doubly robust estimation", "positivity" or "overlap".

[10]The technical term for this compatibility is *congeniality*, coined by Meng (1994), and the concern is practical rather than pedantic. Imputation is often done by one person (or one agency) and analysis by another, and if the imputer's model is poorer than the analyst's (let's say it leaves out interactions, nonlinearities or subgroup structure that the analysis cares about), the completed datasets systematically understate exactly the patterns being studied. The working rule that falls out: make the imputation model at least as rich as the analysis model, and when in doubt, richer. Search "congeniality multiple imputation" or "uncongenial imputation" for discussions.

report a range. Worst-case bounds fill the spaces with the most extreme plausible values in both directions (Horowitz and Manski, 1998), which I think is honest and often uncomfortably wide. Lee trimming narrows the range under one extra assumption (monotone selection), meaning training can move people into employment but never out of it. It also needs the treatment itself to be as good as random, or exogenous given the covariates we use, so it doesn't fix the self-selection into training we worried about in P01. The narrower bound comes at a price in interpretation: it describes the effect for the always-observed group, the people who would be working with or without training[11], and not for everyone originally assigned (Lee, 2009). Fittingly for our example, Lee's original application was a job-training programme, the US Job Corps.

The third response is to reconstruct the missing wages with a selection model. A Heckman-style model writes down who ends up employed and what they earn as one joint system with related errors, and then uses the system to correct the wage equation (Heckman, 1979). For this to rest on more than an assumption about the shape of the error distribution, we need an exclusion restriction[12]: some variable that shifts whether a person works but has no direct business in their wage. Without a convincing one, the identification leans heavily on distributional form, which is a polite way of saying the estimate can move a lot when we change assumptions we can't test.

The fourth response is to stress the missingness assumption directly. Pattern-mixture and tipping-point exercises[13] say: suppose the people we lost differ from the people we kept by some amount, and let's see how large that amount has to be before our conclusion changes (Scharfstein et al., 1999). If it takes a difference that is implausibly large on a scale we can defend, the result is less fragile under that sensitivity model; if a small plausible difference is enough, we've learned our conclusion leans heavily on the missing-outcome assumption (aka, it's held up by hope only).

[11] The formal name for this group is a *principal stratum*, and the idea is worth thirty seconds because it explains who the estimate is really about. Sort people by how their employment would respond to the programme: some would work whether or not they were trained (the always-employed), some would work only if trained, some wouldn't work either way. We can't tell person by person who belongs where since we only ever see each person under one treatment, but the monotone-selection assumption rules out the awkward group that training pushes *out* of work, and that's what lets the trimming produce a bound. The bound then belongs to the always-employed stratum: a policy claim like "training raised wages" silently becomes "training raised wages for the kind of person who'd be employed regardless". The framework is from Frangakis and Rubin (2002). Search terms that work: "principal stratification", "Lee bounds", and the Stata command `leebounds`.

[12] An *exclusion restriction* is a variable that moves the selection (whether we observe the outcome) without moving the outcome itself. In the classic female labour-supply applications, the number of young children was argued to shift whether a woman worked but not her offered wage; whether the exclusion is believable is always the fight, and it should be. What happens without one is specific: the Heckman model is still estimable, because the assumed joint normality of the errors gives the correction term (the inverse Mills ratio) a particular nonlinear shape, and that shape alone identifies the model. That means the estimate is resting on a distributional assumption we can't check, and modest changes to it can move the results a lot. The original is Heckman (1979), the Stata command is `heckman`, and the useful search terms are "Heckman selection model", "exclusion restriction" and "inverse Mills ratio".

[13] A *pattern-mixture model* splits the sample by response pattern, the people we kept and the people we lost, models the ones we can see, and then states openly how the lost ones are assumed to differ, usually through an offset: "dropouts earn 10% less than observationally similar completers". That factorisation by response pattern comes from Little (1993). The *tipping-point* exercise repeats the analysis over a range of offsets and reports the value where the conclusion changes (here "changes" means whatever we declared as the criterion: usually that the effect is no longer statistically distinguishable from zero, sometimes that the estimate changes sign). Instead of defending one assumption, we show the whole range and discuss whether that value is realistic. Scharfstein et al. (1999) is the standard reference for the related selection-model version, where a selection-bias parameter plays the role the offset plays here. One naming warning: this literature (especially in clinical trials where regulators routinely ask for these analyses) calls the offset a "delta adjustment", and that delta has nothing to do with Oster's δ from P01, an unlucky collision of Greek letters. Maybe we should start using Chinese characters instead? :) Search terms: "pattern-mixture model", "tipping point analysis", "delta adjustment", "nonignorable dropout".

So: a reweighting, a set, a reconstruction and a stress test, four different answers to "what happened to the people we can't see?". And we shouldn't expect the data to referee between them because model fit can't do it: every missing-not-at-random model has a missing-at-random counterpart that fits the observed data exactly as well (Molenberghs et al., 2008). The choice between stories is an argument about behaviour, and not a statistic.

P05. A continuous variable is measured with error

Let's change what we're asking of the training study. Suppose we now want to know how earnings gains relate to the hours of training a person actually attended, and hours are self-reported. You see where I'm going with this, right? Nobody remembers precisely; people round to whole weeks and guess. The missing object is the error-free value - or really - the regression relationship we would run if we had it.

Most of us were taught one fact about this situation: noise in a regressor pulls the estimated coefficient towards zero, so our estimate is *at least* conservative. That fact is real but much narrower than its (big) reputation[14]. It holds for classical error, meaning noise unrelated to the true value and to everything else, in a simple linear regression with one mismeasured variable. Add more regressors, let the errors correlate with the truth (remember from P02 that real survey errors do exactly that), make the model nonlinear, or put the error in the outcome instead, and even the direction of the bias can change (Hausman, 2001). "Our estimate is biased towards zero, so the true effect is bigger" is a sentence that needs its assumptions checked before it leaves the seminar room.

The tools for repairing this all work by learning something about the error process, and each one assumes something different about it. A validation sample, as in P02, estimates the error process directly. Repeated measures use a second noisy report of the same variable, and one clean use is as an instrument: the second measure is correlated with the true hours but, if its error is independent of the first measure's error and of the disturbance in our earnings equation, not with the parts that cause the trouble[15] (Griliches and Hausman, 1986). Regression calibration replaces the noisy value with our best estimate of the true one given everything observed, and SIMEX runs the regression several times with extra noise deliberately added, watches how the coefficient degrades, and projects that path backwards to the no-noise case[16] (Cook

[14]The mechanics are worth going over once. A regression coefficient is a covariance divided by a variance. Classical noise in the regressor leaves the covariance with the outcome alone but inflates the regressor's variance, so the ratio shrinks: that's attenuation, and the shrinkage factor is the reliability ratio from P02's footnote, the share of the measure's variance that is signal. The textbook result stops there, and the trouble starts where the textbook stops hehe. With several regressors, the bias from one mismeasured variable spreads to the coefficients of the correctly measured ones, in directions that depend on how everything correlates. If the error is mean-reverting rather than classical, the covariance is contaminated too, and the estimate can be biased away from zero. In nonlinear models there is no single shrinkage factor at all. And error in the *outcome* behaves differently again: if classical, it inflates standard errors without biasing the slope, but if it correlates with the truth (as in the earnings validation studies), it biases the slope as well. Search terms: "attenuation bias", "errors-in-variables", "reliability ratio", and for the multi-regressor case is "measurement error contamination".

[15]We have two noisy reports of the same true variable, say self-reported hours and hours from the training provider's records, which are both imperfect. Using one as an instrument for the other works because a valid instrument needs two things: it must be correlated with the variable we're instrumenting (which may hold if the shared signal is strong enough to give a relevant first stage), and it must be unrelated to the errors that cause the bias (which holds only if the two errors are independent of each other and of the earnings equation's disturbance). That second condition is the one to argue about: if both reports come from the same person's memory, their errors probably share a component, and the instrument inherits the "disease" it was meant to cure. Griliches and Hausman (1986) developed this logic for panel data, where differencing can amplify measurement-error bias, and their paper is also the standard warning about that amplification. Search terms: "instrumental variables measurement error", "repeated measures instrument", "errors in variables panel data".

[16]I'd say these two have the most personality of the correction methods. *Regression calibration* asks: given the noisy report and everything else we observe, what's our best estimate of the person's true hours? It substitutes

and Stefanski, 1994). Likelihood methods write the error process into the model and estimate everything jointly[17].

What all of these can deliver (when their assumptions hold!) is a corrected population parameter: the average relationship between true hours and earnings. What they don't deliver is each person's true hours. The correction works at the level of the regression, and that's usually what we need, but it's worth saying out loud, because a corrected coefficient sitting in a clean table makes it easy to believe the underlying variable was fixed too, and it wasn't.

P06. Either treatment or outcome status is misclassified

Sometimes the mismeasured variable isn't a number but a yes-or-no, and that changes the problem enough to deserve its own entry. In our training study: some people say they attended when they didn't, some attended and don't say so, and on the outcome side "employed" versus "not employed" has its own wrong labels. False positives and false negatives behave differently from the continuous noise of P05 because a binary variable can only be wrong in two directions, and the mix of the two directions is what drives the damage.

The first question is what we're actually trying to recover, because there are three different targets hiding here: the true prevalence (what share of people really attended training), a regression coefficient (how attendance relates to earnings), or a causal effect. Each needs its own correction, and a fix for one is not automatically a fix for the next. A corrected prevalence doesn't by itself correct a regression coefficient, and a corrected coefficient isn't automatically a causal effect (Hausman et al., 1998). This chain is easy to skip in practice: a paper corrects the share and quietly treats the regression as corrected too.

What we can do depends on what we know about the error rates, which readers from epidemiology will recognise as sensitivity and specificity[18]. If we know them, or can estimate them from a validation subsample, a correction can be built for our specific target and error model, and Aigner (1973) shows what that full knowledge buys us in his one-sided model for a binary regressor. If we only have partial information, like a plausible range for the error rates rather than their values, the defensible output is a bound (Bollinger, 1996): a range of coefficients consistent with what we know instead of a corrected point. Between the two sits a third option: when the error rates are uncertain rather than known, probabilistic sensitivity analysis feeds defensible distributions for sensitivity and specificity through the correction and reports the

that estimate (a conditional expectation) into the regression and then adjusts the standard errors to account for the substitution, which makes it simple and popular in epidemiology. *SIMEX*, simulation-extrapolation, is the counterintuitive one: since we can't remove noise, just *add* more of it, in several known amounts, and rerun the regression each time. The coefficient degrades along a path as noise grows, and fitting that path lets us project it back to the point of zero noise (Cook and Stefanski, 1994). The projection step is the assumption: we never observe the zero-noise end, so the answer depends on the fitted shape of the path. Implementations: `simex` in R and Stata's `simex` command; search "regression calibration" and "simulation extrapolation SIMEX".

[17] In the simple classical linear case, we also have two older tools: a known reliability ratio can correct the naive slope directly, and running the regression in both directions (earnings on hours, then hours on earnings, inverted) brackets a positive slope between the two estimates, though neither trick survives nonclassical error or richer models (Hausman, 2001). One caution on regression calibration: it's exact for a linear mean model under its assumptions and generally an approximation in nonlinear ones (Carroll et al., 2006).

[18] *Sensitivity* is the probability that a true yes is recorded as a yes (a real trainee reports attending); *specificity* is the probability that a true no is recorded as a no. Together they pin down the misclassification process for a binary variable, and they make the prevalence correction almost arithmetic: the observed rate of yeses mixes true yeses caught by the test with false positives, so observed rate = sensitivity x true rate + (1 − specificity) x (1 − true rate), and knowing the two error rates lets us solve for the true rate. The same logic extends (with more algebra) to regression coefficients. The reason economists cite Aigner (1973) for the known-rates correction and Bollinger (1996) for the partial-information bounds is that they answered the two different situations: Aigner shows what full knowledge of the error rates buys, Bollinger shows what can still be said when we only know something. Search terms: "sensitivity specificity", "misclassification correction", "misclassified binary regressor".

resulting distribution of corrected results, though it still doesn't learn those distributions from our own data (Fox et al., 2005).

The assumptions doing the work are about the error rates themselves: which ones we know, whether the validation setting is comparable to ours (the transport question from P02 again), and above all whether the rates vary with treatment, outcome or covariates. When they do vary, the misclassification is called differential[19], and it needs its own model, because the convenient folk theorem that misclassification biases us towards zero belongs to the non-differential case, and not always even there.

And when there's no gold standard at all (no validation sample, no error rates from anywhere), latent-class methods[20] can seem to rescue us: give the model several imperfect measures of the same status and let it estimate the error rates and the truth together. The rescue is real but runs on assumptions we should lay out, usually that the measures err independently given the true status and that the classes mean the same thing across populations. The output can look impressively precise, and that precision can be fragile to exactly those assumptions.

P07. A price or quality adjustment is unavailable

Let's forget our training study example for a second because this problem sits in a different corner of economics. Suppose we want to compare the cost of living across regions (spatial variation), or track inflation over time (time variation), and the goods people buy differ in quality (they're heterogeneous): rice in one market is not the rice in another, this year's laptop is not last year's laptop. The missing object may be a comparable price, a quality adjustment, or the weights an index needs. What we observe instead are transactions for goods that are never quite the same twice.

Household surveys tempt us with a shortcut called the unit value: divide what the household spent on rice by the kilos it bought, and call that the price of rice. The trouble is that a unit value mixes three things: the price the household faced, the quality it chose, and the composition of its purchases. One could say that richer households buy better rice, so their unit values are higher even at identical prices, and if we treat unit values as prices we build the household's own choices into the "price" we then use to explain those choices (Deaton, 1988). Unit values can be validated and adjusted into something useful, but they don't start out as prices. Deaton's own method is a reconstruction in this article's sense: it needs geographically clustered households, a model of quality choice and a measurement-error correction, so dividing

[19] Misclassification is *non-differential* when the error rates are the same for everyone regardless of their treatment, outcome or characteristics, and *differential* when they aren't. The distinction decides how much trouble we're in. A classic example from epidemiology is recall bias in case-control studies: people who got sick search their memory for exposures harder than healthy controls do, so the exposure's error rate depends on the outcome, and the bias can go in any direction. In our training example, people who benefited from the programme may be more likely to remember and report attending, which is the same disease. Even the comfortable non-differential case has fine print: the towards-zero result is for a binary exposure in simple settings, and Hausman et al. (1998) show that misclassification in a discrete *outcome* biases coefficients in probit and logit models even when the errors are non-differential, which is the nonlinear-model warning from P05 wearing a different coat. Search terms: "differential misclassification", "non-differential misclassification bias", "recall bias".

[20] The best-known design is due to Hui and Walter: with two imperfect tests applied in two populations whose true prevalences differ, the error rates of both tests and both prevalences become estimable at once, with no gold standard anywhere. It feels like magic, and the magic has a price list. The measures must err independently given the true status, which is exactly what stops holding when both measures come from the same interview or the same administrative pipeline, and the true classes must be the same objects across the populations, and the design adds two requirements of its own: the two populations need different true prevalences, and each test's sensitivity and specificity must stay constant across them (Hui and Walter, 1980). When any of these is wrong, the model may still converge and print tight standard errors (Albert and Dodd, 2004), which is why an apparently precise answer can be fragile: the precision is conditional on assumptions the data can't check. Search terms: "latent class analysis without gold standard", "Hui-Walter design", "conditional dependence diagnostic tests".

expenditure by quantity is where the exercise starts rather than what it is.

For goods that change quality over time, hedonic methods[21] price the characteristics instead of the good: regress observed prices on attributes (for a laptop, memory, speed, screen), and use the fitted relationship to ask what the new model would have cost with the old characteristics (Rosen, 1974). The comparison this gives us depends on the characteristics we observed and the functional form we chose, which is why official price indices don't stop at a regression: the statistical manuals spell out explicit rules for matching items over time, weighting them, handling a product's replacement and linking new items in (Intersecretariat Working Group on Price Statistics, 2020). The rules don't remove the judgement, but they put it where we can inspect it. Two named index families do much of this work: matched-model indexes compare identical items over time, and multilateral methods pool comparisons across several periods, which also reduces the chain drift we will meet below, though both still depend on the matching rule, the product universe and how entry and exit are handled (de Haan and van der Grient, 2011).

There's also a difference between calculating a missing price and inferring one, and it decides how much the number can be trusted. Where a fully specified accounting identity pins the number down, total expenditure, quantities and the other components all observed and compatible, we can calculate the residual and carry the components' uncertainty through to it. A price backed out of a demand curve, a production function or any other behavioural relationship is a different object: a model-based scenario whose value changes with the model. There is no general "residual price" correction. And the diagnostics here have the usual limit. Chain-drift checks[22] can tell us a chained index has "wandered" in a way that pure price change can't explain, but they can't tell us whether the items being linked were economically comparable in the first place.

P08. The counterfactual is unavailable

This is THE central absence in causal inference, the one every other previous entry has been circling around. Let's take our example a level up: a state introduces a training subsidy and we want to know what happened to earnings *because* of it. The missing object is what earnings in that state would have been without the subsidy. No dataset ever contains that column so everything in this entry is about constructing a stand-in for it and then interrogating (a lot) the construction.

[21] The hedonic idea is that a differentiated good is a bundle of characteristics, and the market implicitly prices the characteristics. Practically it's a regression of price on attributes, and statistical agencies use exactly this to quality-adjust fast-changing goods like computers and housing. Rosen (1974) is the theory paper behind the practice, and it comes with a warning that often gets lost: the regression coefficients are equilibrium implicit prices, formed where buyers' preferences meet sellers' costs, and not demand parameters or willingness to pay. Two things limit any hedonic adjustment: the characteristics we didn't record (a camera improves in ways no spec sheet captures) and the functional form we imposed. Search terms: "hedonic regression", "hedonic quality adjustment", "implicit prices".

[22] A chained index links many short comparisons: it compares this month with last month, then multiplies that by the comparison of last month with the month before, and so on. Chain drift is when this multiplication goes wrong in a specific way: prices and quantities come back exactly to where they started, but the index doesn't. One well-documented cause is supermarket sales in scanner data (de Haan and van der Grient, 2011). During a sale the price is low and households buy a lot, so the price cut enters the chain with a big weight. When the sale ends the price goes back up, but by then purchases have returned to normal, so the price increase enters with a smaller weight. The cut and the increase don't cancel out, the index ends the round trip below where it began, and over many sales the gap keeps growing. A chain-drift check tells us this is happening, and the multilateral methods built to avoid it can reduce the drift under a stated product universe, window and formula. What they can't answer is whether a discounted, end-of-line item is still the same good as its full-price predecessor: that is a judgement about comparability, and - sadly - no statistic settles it. Search terms: "chain drift", "scanner data price index", "GEKS".

Design, diagnostics, falsification and sensitivity analysis constantly get blurred under the word "robustness", and it helps to keep the four apart. Our chosen design constructs the comparison: these units, these periods, this assumption about why the comparison is fair. The diagnostics check parts of that design, while falsification tests look for effects in places where the design says there should be none, and sensitivity analysis varies an identifying restriction to report which conclusions remain. The four are related, and they are not the same task, so a paper that runs one of them has not automatically done the others. For our subsidy question, the two designs we would most likely use are synthetic control and difference-in-differences[23].

With synthetic control, we build the counterfactual ourselves: the untreated states get combined into a weighted mix, and the weights are chosen so that the mix tracks the treated state's earnings path before the subsidy (Abadie et al., 2010). The comparison is only as good as its conditions, good pre-treatment fit and a donor pool of comparable, untreated states. The usual check is a placebo run[24]: pretend each donor state passed the policy and see whether our real state's post-policy gap stands out among the pretend ones. That comparison produces a rank, and a rank is not a p-value by itself (Abadie, 2021); it earns a probability interpretation only with an argument about how treatment was assigned. There's also an in-time placebo: move the intervention to a date before the real one and check whether the same procedure finds an effect where there shouldn't be one, and here a check that finds one counts against the design, while a clean one stays just a diagnostic (Abadie, 2021).

With difference-in-differences, the counterfactual comes from assumptions rather than weights, and two of them work together. They are a) parallel trends, which in our example would be "without the subsidy, the treated state's earnings would have moved like the comparison states' did", and b) no anticipation, in our example "earnings didn't start responding before the subsidy took effect because people saw it coming". *(Anticipation and treatment timing get their own entry in upcoming P19; here we stay with parallel trends, which is where - one can say - the diagnostic action is - wink).* The standard diagnostic is a pre-trend check, and it's a bit weaker than its popularity suggests[25]: the test can have little power against the violations that would hurt us most, so passing it doesn't validate the assumption (Roth, 2022). The newer response is to stop testing and start stressing: the Honest DiD approach[26] asks what we can

[23] As readers of DiDDigest can remember, both designs have newer hybrids: augmented synthetic control brings in outcome modelling to patch residual pre-treatment imbalance (Ben-Michael et al., 2021), and synthetic difference-in-differences combines unit and time weighting with a DiD adjustment (Arkhangelsky et al., 2021).

[24] A step by step of the placebo logic: run the entire synthetic-control procedure once per donor state, each time pretending that donor passed the policy, and collect the resulting post-"policy" gaps. If our real state's gap is the largest of twenty (or something) such runs, we can say its rank is 1 out of 20, and it's tempting to call that $p = 0.05$, but we should be really careful because that number is a real randomisation p-value only if the policy was as good as randomly assigned among the states, and nothing in the procedure makes that true since states pass subsidies for n different reasons. Without that argument, the rank is still informative, our state stands out, but it's a comparative statement instead of a significance level. Abadie (2021) is the survey treatment of this and of the feasibility conditions, and it's the right first read before using the method. Search terms: "synthetic control method", "placebo test synthetic control", "in-space placebo"; implementations are `synth` (Stata, R) and `tidysynth` (R).

[25] A pre-trend check regresses the outcome on time relative to treatment and asks whether the treated and comparison groups were already diverging before the policy. Roth (2022) documents two problems with it. The first one is low power: when the data are noisy, a violation large enough to bias our estimate badly can still pass the test, so a finding of "no significant pre-trend" often means we couldn't detect the divergence rather than that there was none. The second problem comes from the pre-testing itself: if we only carry on with the analysis when the pre-test looks clean, then the estimates that end up published are a selected sample, and that selection introduces its own bias. None of this makes pre-trend plots useless since they still show what the two groups did before the policy; the trouble starts when we treat a passed test as a validated assumption. Search terms: "pre-trends test power", "pretesting parallel trends", and Prof Roth's `pretrends` R package.

[26] The Honest DiD idea is to replace "parallel trends holds" with "parallel trends can be violated, but only by this much", and then to state the "this much" as a number. One version allows the post-treatment violation to be at most M times the largest pre-treatment violation, so $M = 1$ means the assumption can be broken after the policy by as much as it ever was before. Another version restricts how fast the violation can change from

conclude if post-treatment violations of parallel trends are allowed to be, say, no larger than the worst pre-treatment one, and reports an interval or a set for the effect under that restriction (Rambachan and Roth, 2023). The restriction, and not the method's reassuring name, is what does the identifying work.

P09. Inputs or computational objects are inaccessible

This entry is about a different kind of disappearance, and the example works best told forward. Ten years from now, someone questions our training-subsidy paper. The tax microdata we used were confidential and our access agreement has expired, the person who wrote the estimation code has left, and what survives is the published paper with its tables (this hits way too close to home). The missing objects are the inputs of the analysis itself: microdata, code, the fitted model, the intermediate files that turned one into the other. Every other entry in this guide worried about data on the world; this one is about the data trail of the research.

The first thing to keep in mind is what question each exercise answers here, because the vocabulary gets used loosely[27]. Reproduction asks whether the same data and the same instructions regenerate the published number (Dewald et al., 1986; Chang and Li, 2022). That is a question about the record, and it's separate from whether the design identifies anything, from whether the finding is true about the world, and from whether a new study with new data would find it again. A number can reproduce and still be wrong, and a paper whose files were lost can still have been right. We could never know.

Some questions need more than the published record to answer. If we want to know whether one state, or a handful of unusual people, drove our earnings estimate, we need influence or leave-one-out calculations[28], and those run on the objects themselves: unit-level contributions, the fitted model, score contributions, or something equivalent that survived. When those objects are gone, they are gone. The defensible report in that situation is a documented limit on reproducibility, saying what was checked, what could not be, and why; an influence analysis

period to period (a smoothness restriction). For each choice of restriction, the method reports the range of effects consistent with the data, and the usual summary is the breakdown value: the M at which our conclusion changes, which in most applications means the M where the reported interval starts to include zero (Rambachan and Roth, 2023). Reading a result then becomes an argument about whether violations of that size are plausible in our application, and that argument is on econ rather than on stats. The implementation is the `HonestDiD` package (R and Stata); search terms: "Honest DiD", "Rambachan Roth", "relative magnitudes restriction", "breakdown value".

[27] Discussing the taxonomy is important here since the words get swapped quite frequently in seminars: *reproduction* (or computational reproducibility) means running the same data through the same code and getting the same number; *replication* means asking the question again with new data or a new design; and both are different from *identification* (does the design isolate the effect?) and from empirical validity (is the claim true about the world?). The reason economists cite Dewald et al. (1986) is the story: the *Journal of Money, Credit and Banking* asked authors for their data and programs, and most published results could not be regenerated, often for silly reasons like lost files and undocumented steps. Chang and Li (2022) repeated the exercise decades later (sixty papers from thirteen journals) and their title answers the question: "often not". The silly causes are the point of this entry: in those studies, results became irreproducible through ordinary decay, lost files, expiring agreements and departing coauthors, which doesn't rule out misconduct elsewhere, but does show how far ordinary entropy gets on its own. Search terms: "computational reproducibility economics", "replication in economics", "data availability policy".

[28] An influence analysis asks how much each observation contributed to the final estimate, and a leave-one-out analysis answers it by brute force: drop one unit, re-estimate, repeat. Both need - at least - the per-unit material, the microdata or the fitted model's unit-level pieces (in regression these are the score contributions, each observation's push on the coefficient). This is why they can't be run from a published table: a coefficient and a standard error are sums over units, and a sum doesn't remember its parts. The practical lesson runs in both directions. Looking backward, if the objects are gone, the analysis is gone. Looking forward, archiving the fitted model and unit contributions alongside the code is what keeps this door open for our own papers. Search terms: "influence function", "leave-one-out", "dfbeta", and for the recent automated version "finite-sample robustness metric".

can't be recreated from objects that no longer exist, and writing one up as if it could is a reconstruction wearing a "diagnostic's clothes".

Before settling for that report, there are moves we haven't used yet, and they're the same verbs from the beginning of this article. The first one is to reconstruct the inputs from their sources instead of from the paper: the tax microdata still exist at the agency even though our agreement expired, agreements can be renewed, and if we documented our extraction criteria, a new extract plus a reimplementation of the described pipeline can regenerate the analysis. The result needs plain labelling, because it's a new extract of possibly revised data (this is P22's problem, arriving a bit early), so a difference from the published number can come from the data as well as from the code.

The published record itself can also be checked without touching any lost object, because a table has arithmetic obligations: the coefficients have to cohere with their standard errors and t-statistics, the sample sizes have to agree across tables, and the subgroups have to add up to their totals. This is diagnosis rather than regeneration: it can reveal contradictions in the surviving report, and it doesn't regenerate the missing analysis (psychology built automated versions of these checks; search "statcheck" and "GRIM test").

There is a narrow path between giving up (and crying on the floor) and pretending. On that path, a scenario range or a set of bounds can stand in for the lost analysis, but only after we define which missing inputs we consider compatible with the published record and justify that set: "the result holds under any reasonable version of the lost code" is an empty sentence until "reasonable" has content. Even then, the published totals on their own rarely settle it because a table of coefficients doesn't reveal how each unavailable component, a weight here, a sample restriction there, moved the original estimate. When none of this recovers the analysis, the remaining move is to change the question: a replication with new data stops asking whether the published number was computed correctly and asks whether the claim is true, which is usually the thing we cared about in the first place.

P10. Only aggregates are observed

Suppose the individual data behind our training question were never available to us in the first place, and all we have is a table by state: the share of workers who went through training and the average earnings. The missing object is the joint distribution, which is the individual relationship living inside those margins. The temptation we face is to correlate the two columns and say that trained workers earn more, and this temptation has a name and a famous cautionary tale[29]: the correlation between two aggregates can differ in size - and even in sign - from the correlation between the same two things across people.

Once we stop trusting the two-column correlation, the correct move is to ask less of the aggregates we have, and ecological bounds are the exercise built for that: they describe the set of individual relationships consistent with the margins we observe, and the set is often wide, because the margins don't determine what happens inside them (Cho and Manski (2008); open-access chapter here). Sadly, the bounds don't recover the micro relationship, and more aggregate margins alone won't either; a narrower set or a point estimate comes only from additional restrictions, and then the restrictions carry the conclusion. King's ecological-inference model is

[29]The tale is Robinson (1950), and it's worth knowing in its original form. Across US states in 1930, the share of foreign-born residents correlated *positively* with English literacy rates: states with more immigrants had higher literacy. Across individuals, the correlation ran the other way since immigrants were on average less literate in English than the native-born. Both facts are true at once because immigrants had settled in states with high native literacy. Reading the state-level correlation as a statement about individuals became known as the *ecological fallacy*, and Robinson's paper is the reason every methods course warns about it. Search terms: "ecological fallacy", "ecological inference", "cross-level inference".

the best-known point-estimation route, and it works exactly that way: it selects an answer from what the margins allow by adding distributional and homogeneity assumptions that the margins themselves can't test (King, 1997). The available diagnostics tell us more about the aggregation than about individuals: re-estimating across plausible geographies (like counties instead of states, if possible) shows how sensitive the aggregate association is to where the (sometimes literal) lines were drawn, a dependence with its own name (the modifiable areal unit problem, or MAUP), and comparing the aggregate analysis with an individual-level one, wherever even a small individual sample exists, can expose the contextual and ecological biases in the aggregate estimate[30] (Greenland, 2001).

Multilevel and within-between models sound like the modern solution, and they are, but only for the data they actually need: relevant lower-level observations, or repeated hierarchical information such as many periods per state. Given one aggregate cross-section, a multilevel model has nothing to work with underneath the aggregates; it can't manufacture individual data from a table of forty-something rows.

Shift-share and component decompositions[31] can also mislead here, because they look like they answer the individual question and they don't. A decomposition tells us how the aggregate change adds up, how much came from composition shifting across groups and how much from changes within groups, and that's a descriptive answer to a descriptive question. It isn't a solution to ecological identification, because both pieces of the sum are themselves aggregates.

The useful discipline is to state which margins we observe, which associations are missing, and what extra structure would be needed to connect the two. Papers that do this read differently, because the reader can see where the aggregates stop and the assumptions start.

P11. Definitions or classifications change

Going back to our training study for this. Suppose we can now track earnings by occupation for twenty years and halfway through the occupation classification gets revised, e.g. "web developer" didn't exist in the old system, "typist" is practically gone in the new one, and a category like "clerical worker" was split into others that now occupy different places. The missing object is comparability - over time or across systems. Nothing was mismeasured and nobody dropped out, but the ruler itself changed. Readers from other fields will recognise the situation because it's the same one that affects trade data when product codes are revised and health data when ICD versions turn over: it's the problem from P07, where this year's laptop was not last year's laptop, except now it's the categories that are never quite the same twice (Heraclitus vibes).

We have three direct ways to connect the classifications: a crosswalk, a splice built from an overlap, and a backcast. Chain-linking is a related time-series operation: it joins adjacent price

[30] The aggregate association mixes up to three things, and Greenland (2001) is the careful accounting of them. There is the individual effect we wanted (does training raise your earnings?), a possible contextual effect (does living in a high-training state raise your earnings through the labour market, whatever you did yourself?), and confounding at the group level (high-training states differ in other ways). The contextual effect deserves respect rather than deletion because sometimes it's the actual question since spillovers from a trained workforce are a real policy object; the fallacy here is in reading an aggregate coefficient as an individual one without saying which of the three ingredients it contains. Search terms: "contextual effects", "ecological bias", "multilevel analysis".

[31] A shift-share decomposition splits an aggregate change into parts using an accounting identity: the change in a state's average earnings, for example, into the part from employment shifting between industries (the composition part) and the part from earnings changing within industries. The arithmetic is exact and the output is descriptive. One naming warning because economics reused the phrase: "shift-share" also names the Bartik instrument literature where industry composition is used as an identification strategy, and that is a different object with its own assumptions rather than an accounting identity. A reader searching should use "shift-share decomposition" for this footnote's tool and "shift-share instrument" or "Bartik instrument" for the other one. Related decompositions worth knowing by name: "within-between decomposition" and "Oaxaca-Blinder".

or volume comparisons, but it doesn't by itself translate an old category into a new one. A crosswalk[32] builds the bridge from concordance information, meaning documented links between the two systems (Pierce and Schott, 2012). A splice builds it from an overlap period - a stretch where both systems ran side by side - so we can see directly how they line up. Backcasting applies an existing bridge backwards, rewriting the earlier data in the new system's terms, so the clerical workers of 1985 get re-expressed as today's categories. And chain linking replaces one long bridge with many short ones: it compares each year only with the next one, inside whichever system covers both, and multiplies those short comparisons into one long series (International Monetary Fund, 2018). The shared assumption is the same for all the first three: a bridge is still meaningful only as far as the observations used to build it, and the employment shares we measured in the overlap years, applied to 1985, assume the occupation's composition hadn't changed by then, which is exactly the kind of thing that does change.

These exercises can produce a usable series, but no bridge can erase a substantive change in what a category means. If the new "software occupations" group covers different work than the old "computer operators" did, a perfectly smooth spliced series is describing a category whose meaning moved beyond it, and the smoothness is the disguise rather than the reassurance. What we then need to do is to disclose clearly what we did since we can't expect the reader to audit a bridged series without knowing the many-to-many mappings, the break dates, the overlap years, and how entrants and exits were handled. What helps is to create a revision triangle[33], which describes how a statistic's estimates change across successive releases, with one row per reference period and one column per release date, so reading along a row shows how the estimate for that period changed as new information arrived. It's a diagnostic for how the data arrive, but it has nothing to say about a classification break. Revisions are their own absence with their own exercises (which we will see in a second).

P12. The sample is small or the event is rare

Suppose our training programme was only a pilot: six sites, two hundred people, and the outcome we care most about - moving into self-employment - happened eleven times. Two separate problems come from this situation, and they need different tools to be handled. The first is that p-values and confidence intervals depend on large-sample approximations, and with six sites and eleven events, the reference distribution behind them can be a poor description of the uncertainty we face in reality. The second is that the model itself can become unstable or separated[34], and with something like eleven events we are certainly in that territory.

[32] A concordance is the documented mapping between two classification systems, and the issue here is that it's rarely one-to-one in the sense of one old category splitting into several new ones, and several old ones merging into one new one. Allocating a many-to-many mapping needs weights (something like employment, output or trade shares from the overlap period), and those weights are where the assumptions enter, since a share measured in 1997 is being asked to describe 1985. Pierce and Schott (2012) built the standard concordance between ten-digit US trade codes and industry classifications and their paper kind of lists everything that can go wrong. The same structure exists in other fields under other names: medicine crosswalks ICD-9 to ICD-10 with official equivalence mappings, and education does it between ISCED versions. Search terms: "concordance", "crosswalk", "classification change", and for the trade version search "Pierce Schott concordance".

[33] A revision triangle is a table with one row per reference period (e.g., GDP in 2019Q4) and one column per release date, so reading along a row tells you how the estimate for that quarter changed as later releases arrived. It's called a triangle because recent periods have had fewer releases so the table has a staircase edge. Agencies and real-time-data archives publish these, and they answer questions like "how large is a typical revision?" and "do first releases systematically under- or overstate growth?". What a triangle can't do is repair a classification break because a break is a change in what was being counted, and it shows up in the triangle as a discontinuity that no amount of averaging across releases eliminates. Search terms: "revision triangle", "real-time data", "data vintages".

[34] Separation is the situation where our predictors split the two outcome groups perfectly: you could draw a line with every self-employment case on one side and every non-case on the other, e.g. all eleven cases sitting in trained sites and none anywhere else, so the variable "trained site" sorts the yes's from the no's on its own.

For the first problem we have three standard solutions - randomisation inference, permutation tests, and the wild-cluster bootstrap[35] - and each one builds its comparison world from a different source. Randomisation inference uses the assignment mechanism we are familiar with: if we know exactly how the six sites were chosen for treatment, we can then re-run that assignment thousands of times (on our software of choice), and under the null of no effect for anyone, we can fill in what each site would have shown so the p-value counts how unusual our real allocation looks among all the allocations that could have happened (Athey and Imbens, 2017). A permutation test shuffles labels instead, and it needs an exchangeability argument - a reason to believe the shuffled versions are statistically interchangeable with the real one under the null (Chung and Romano, 2013). The wild-cluster bootstrap resamples the data in a particular way, and the quality of what it produces depends on the details we have: how much leverage each cluster has, how treatment is spread across clusters, whether the null is imposed when resampling, the choice of weights, and the statistic being bootstrapped. Their shared commonality is that each of these builds its reference world from one source, a known mechanism, an invariance argument, or a resampling construction, and the result is only as good as that source. There's also a fourth, albeit less glamorous, family that adjusts the cluster-robust variance and the degrees of freedom directly for having few clusters (Cameron and Miller, 2015). These are corrections to an approximation rather than new reference worlds, and how well they do still depends on leverage, on balance, and on how many clusters actually got treated.

For the second problem the tools work on the estimate rather than on the reference world - Firth correction, rare-events corrections, and Bayesian partial pooling. Firth correction is a penalised version of maximum likelihood that reduces small-sample bias and keeps the estimate finite when the data are separated (Firth, 1993)[36], whereas rare-events corrections deal with the small-sample bias of logistic models when the outcome is uncommon (King and Zeng, 2001). The right correction depends on why the events are rare and on how the sample was drawn, and recovering the population probabilities can require extra information like the population prevalence or the sampling fractions. Bayesian partial pooling shrinks noisy site-level estimates towards each other, trading them against a hierarchical model and a prior (Gelman, 2006). All three estimate or regularise, and that solves the second problem only. The first one is still standing: before we can report a p-value or a confidence interval, we still need a reference distribution to compare the estimate against, so the tools in this paragraph and the tools in the previous one do different jobs, and a small-sample study often needs one from each.

Even the best reference world is still small here. With six sites split three against three, there are only twenty ways treatment could have been assigned, so the smallest p-value randomisation inference can produce is $1/20 = 0.05$, and that's for a one-sided test where our real allocation is the most extreme of them all. With the usual two-sided statistic the floor is generally $2/20 = 0.10$ because every allocation has a mirror allocation with the opposite sign. A placebo-

Ordinary maximum likelihood then has no finite best answer because making the coefficient larger always fits a little better, so the estimate runs off towards infinity, the standard errors explode, and some software stops with a warning while some reports enormous numbers. "Unstable" means that the estimate stays finite but swings considerably when one or two observations change. The fixes are in the main text (Firth correction, rare-events adjustments); the implementations are `logistf` in R, `firthlogit` in Stata, and `relogit` for the King–Zeng version. Search terms: "separation logistic regression", "complete separation", "quasi-separation".

[35] Cameron et al. (2008) found substantial improvements in simulations with as few as five to ten clusters, though that's no general guarantee because with treatment concentrated in one or two clusters, or badly unbalanced cluster sizes, the bootstrap can still perform poorly. The practitioner's guide to all of it is Cameron and Miller (2015). One configuration is tough even for the bootstrap: few treated clusters, e.g. six sites of which only one or two got the programme, where no resampling scheme has much to work with and the randomisation-inference route tends to be the more defensible one. The implementation everyone uses is `boottest` in Stata (also available in R); search terms: "wild cluster bootstrap", "few treated clusters", "cluster robust inference".

[36] To be fully correct, Firth's bias-reduction adjustment modifies the likelihood score (Firth, 1993), and its use to obtain finite logistic estimates under separation was established by Heinze and Schemper (2002).

style rank deals with the same arithmetic: standing out among a handful of alternatives is a comparison, and it becomes a formal p-value only with an assignment or invariance argument behind it. None of the tools in this entry escapes this, and none of them repairs a weak design, poor measurement, or a target population the sample doesn't support. What they do is make the most of six sites while being clear about how much that is - and with six sites, saying which population our estimate speaks for is often the hardest sentence in a study.

P13. There is little overlap or little identifying variation

Back to the observational version of our training study where people chose whether to enrol. Suppose the choosing was "lopsided": almost everyone with a university degree signed up, and almost nobody over sixty did. For those two groups the data contain no comparison because there are no untreated graduates and no treated sixty-somethings to compare against, and no amount of regression adjustment would conjure that. The missing object is the support for the original comparison: treated and untreated people who resemble each other across the whole range of characteristics our question covers.

We have two standard responses - trimming and overlap weighting[37] - and they share one consequence: both change who the estimate is about. Trimming is about dropping the observations where a comparison is unsupported, e.g. the graduates and the over-sixties, and then estimating the effect for the people who remain (Crump et al., 2009). Overlap weighting keeps everyone but counts people more the more "contested" their enrolment was so the estimate concentrates on the middle ground where similar people ended up on both sides (Li et al., 2018). Neither recovers the effect for the full original population, and the recommendation that follows is the same as in P11: explicitly write/say what the new estimand is and who was excluded since "the effect of training" and "the effect of training for people whose enrolment could have gone either way" are very different sentences.

Little identifying variation is the same absence in IV form. Suppose we use distance to the nearest training centre as an instrument for enrolment, and distance turns out to barely move anyone's decision: the first stage (the relationship between the instrument and the treatment) is then fragile, and everything built on it inherits the fragility. The diagnostics, like the familiar first-stage F statistic, describe that fragility without fixing it (Staiger and Stock, 1997). Weak-instrument-robust procedures[38], like the Anderson-Rubin test, change how the uncertainty is computed so that the reported intervals stay valid however weak the first stage is, and that is all they change: they don't strengthen the instrument, and they don't validate it (Andrews et al., 2019). Their guarantees also come from a specific setting, and the classical exactness

[37] Both tools run on the propensity score, which is the estimated probability of enrolling given a person's observed characteristics. A score near 0 or 1 flags a person whose treatment status was close to decided in advance, and this is exactly where comparisons are unsupported. The practical trimming rule from Crump et al. (2009) is to keep observations with scores between roughly 0.1 and 0.9, and it's a rule of thumb rather than a law. Overlap weights, from Li et al. (2018), weight each person by the probability of the treatment they didn't get, so the weight peaks for people with scores near 0.5 and fades toward the extremes, and the resulting estimand even has its own name, the average treatment effect in the overlap population (ATO). Search terms: "common support", "propensity score trimming", "overlap weights", "limited overlap".

[38] The famous $F > 10$ rule of thumb for the first stage comes from Staiger and Stock (1997), and it was a rough guide from the start; the modern survey treatment of what to do instead is Andrews et al. (2019). The robust procedures work by inverting tests rather than trusting the usual standard errors: the Anderson-Rubin test asks, for each candidate effect size, whether the data reject it, and collects the survivors into a confidence set. One feature of these sets is worth knowing before you meet one: when the instrument is weak enough, the set can be wide or even unbounded, and that is the method being truthful about the data rather than the method misbehaving. Anderson-Rubin also has company: conditional likelihood ratio tests can say more in the settings they cover, and the first-stage F we all learned to check has to match the covariance structure of the data, so the usual threshold is not a certificate once heteroskedasticity, clustering or serial dependence enter the picture (Andrews et al., 2019). The implementations are `weakiv` and `ivreg2` in Stata and `ivmodel` in R. Search terms: "weak instruments", "Anderson-Rubin test", "conditional likelihood ratio test", "first-stage F statistic".

results don't carry over on their own to clustered or serially dependent data, so the version we run has to match the dependence structure we have.

The two halves of this entry are the same absence in two forms: the data hold little information where our question needs it. The defensible responses either shrink the question to where the information is, and are explicit about it, or keep the question and report uncertainty in a way that respects how little there is. Nothing here manufactures variation, and an estimate that ignores that will look precise for the wrong reason.

P14. Records describing the same units can't be linked exactly

Our training survey has two hundred people, and the tax records that would tell us their true earnings sit in another file with no shared identifier. Linking means deciding which tax record belongs to which respondent using name, birth date and address, and all three can disagree for the same person: names get typed differently, people move houses, and some change their name at marriage. The missing object is unit identity across files - the fact of the matter about who is who.

Probabilistic linkage[39] turns this decision into a calculation. Each candidate pair of records gets compared field by field, the pattern of agreements and disagreements goes into a comparison model, and the model returns a match probability[40] - a pair that agrees on birth date and address but differs on one letter of the surname is probably the same person, and a pair that agrees only on being called "J. Silva" probably isn't (Fellegi and Sunter, 1969). For this to work we need three things: 1. linking fields that carry information, meaning fields that distinguish people rather than describe half the country, 2. a comparison model we can defend, and 3. a manageable set of candidate pairs, because comparing two hundred respondents against every tax record in the country is neither feasible nor necessary.

The linkage will make two kinds of error, and they do different levels of damage. A false link attaches someone else's tax record to our respondent, so their "true earnings" are now a stranger's. A missed match loses the record that was there to be found, and the people it loses are not random since the hardest people to link are the ones who - in our example - have moved, divorced or changed names. What this does to our estimate depends on the full structure of those errors, and it need not be a simple attenuation (Lahiri and Larsen, 2005), so the defensible practice is to carry the linkage uncertainty into the analysis, either by keeping the match probabilities and letting them propagate, or by modelling the linkage and the analysis as one joint problem. And when the files can only match one-to-one (each person has at most one tax record), the candidate links stop being independent choices, so the model should enforce that structure and the joint versions can then average the analysis over several complete, mutually compatible linkages.

The diagnostics assess the process, and the process is not the population. Clerical review (humans checking a sample of borderline pairs) and quality metrics like match rates tell us how the linkage went for the records we tried to link. A high match score doesn't establish

[39] The Fellegi-Sunter framework runs on two probabilities per linking field: how often the field agrees when two records truly are the same person (high, but not 1, because of typos), and how often it agrees by coincidence when they aren't (low for birth dates, higher for common names). Each field's agreement or disagreement contributes evidence, the contributions add up to a score, and two thresholds turn scores into decisions: clear matches above, clear non-matches below, and a middle zone sent to clerical review. The "manageable candidate pairs" requirement has its own name: blocking. Rather than comparing everyone with everyone, we compare only within blocks that share something cheap and reliable, like a birth year, and accept that a wrong birth year now means a missed match. Implementations: `fastLink` in R, `dtalink` in Stata, and `Splink` in Python. Search terms: "probabilistic record linkage", "Fellegi-Sunter", "blocking record linkage", "linkage error bias".

[40] Some later models turn that evidence into an actual match probability, but only by adding assumptions about the linkage process (Sadinle, 2017).

that the linked sample represents the target population: if the people we failed to link are the movers and the name-changers, our linked sample is simply a sample of stable lives, and that is a selection problem in a data-management costume.

P15. Variables sit in different datasets

This is the sibling of P14 with one twist now: the files describe different people. Our training survey has enrolment and hours for two hundred respondents, and earnings are in a labour-force survey of other people entirely, so there is no pair of records to link, no matter how clever we get. The missing object is the cross-file association: nobody, anywhere, observed training hours and earnings together, and that association is exactly what our question needs.

One way out is to give every respondent a statistical twin. We look in the other file for someone with the same age, education and from the same region, we copy that person's earnings over, and we end up with a completed dataset where everyone has both variables. This exercise is called statistical matching[41], and the donated association comes from an assumption, usually conditional independence: given the shared variables, hours and earnings are assumed unrelated, so the shared variables are assumed to carry all the connection there is (Rässler, 2004). The completed file can't test this because the association it contains is the one the assumption put there. A completed synthetic microdata file is evidence that an algorithm worked, but we can't say that the missing association was recovered.

The other way out skips completed files: instead of building records that have both variables, we take summary pieces from each file (like the relationship between distance and enrolment from one survey and between distance and earnings from the other) and combine the pieces into the estimate we want. The exercises in this family are moment combination, two-sample IV, and two-sample 2SLS[42], and they are separate exercises: each uses different pieces from each file, and each needs its own conditions (Ridder and Moffitt, 2007; Inoue and Solon, 2010). The shared requirements are a) compatible variable definitions, b) the same population relationships holding in both samples, and c) variance conditions that differ by estimator - and these sit on top of the ordinary IV requirements[43]. When those conditions ask more than we can defend, we can report a set instead - a range of answers consistent with what each file separately shows, instead of a fused point estimate carrying an unverified association inside it.

[41] The mechanics are usually a donor pool and a distance rule: the other file's records are the donors, we measure similarity on the shared variables, and the closest donor's value gets copied over (or several donors get averaged, or a model predicts the value from the shared variables - the flavours differ, the logic doesn't). The name for what's being assumed is the conditional independence assumption, and we can see how it works in practice in our example: if motivated people train more hours and earn more in ways age, education and region don't capture, the matched file will understate the hours-earnings connection, and nothing in the file will show it. Without conditional independence, the data from the two files only bound the association rather than pin it down, which is why the fusion literature increasingly reports uncertainty about the match itself. Search terms: "statistical matching", "data fusion", "conditional independence assumption", and for the bounds version the famous "Fréchet bounds".

[42] Two-sample IV runs an IV design with its two halves in different datasets: one file has the instrument and the treatment, so it can estimate the first stage, and the other has the instrument and the outcome, so it can estimate the reduced form, and the ratio of the two estimates the effect - under the usual IV assumptions - even though no single person has all three variables. In our example, one survey shows how distance to a training centre moves enrolment, the other shows how distance relates to earnings. The requirements are that both samples come from the same population and that the variables mean the same thing in both, which sounds ok and reasonable until two surveys define "earnings" differently. Drawing both samples from the same population is the cleanest case; there are methods that let the samples differ (Zhao et al., 2019), but they work by stating which relationships stay stable across the two, not by waving the difference away. Two-sample IV and two-sample 2SLS are not the same estimator: Inoue and Solon (2010) show they differ in finite samples and in efficiency, so you have to choose between them. Search terms: "two-sample IV", "two-sample 2SLS", "data combination econometrics".

[43] Splitting the moments across two files doesn't excuse the instrument from its "day job": it still has to move treatment, affect the outcome only through treatment, and be unrelated to the unobserved determinants of earnings.

Whatever we choose, the useful discipline is to name - again - the bridge: say whether the information is being combined across files, across instruments, or across populations because each link has its own weak point. For a statistical match it is the conditional-independence assumption; for a two-sample IV it is two samples that don't come from the same population; and neither problem is revealed automatically in the output, which looks complete either way.

P16. Values are censored, top-coded or altered for disclosure

One more visit to our training survey :) This time the data arrive incomplete on purpose. Earnings above 150,000 show up as "150,000+", some respondents answered with a bracket ("between 30,000 and 40,000") - not their fault! - and in the public state tables, cells with too few people are left blank. The missing object depends on the release mechanism (the rule the publisher applied before letting the data out), and the mechanisms are a family with true differences[44]: a top-code replaces values above a limit with the limit itself, a bracket gives a range instead of a number, censoring cuts off at a known point, swapping exchanges values between similar records, rounding coarsens them, noise addition perturbs them, and suppression removes cells outright. Lots of problems here because each of these actions throws away different information, therefore... yeah, you guessed, each asks for a different exercise.

For the two oldest mechanisms, we can name the exercises. When the data give us ranges, we can fit a regression that uses the known endpoints of each range instead of a single value, and that is known as interval regression (Stewart, 1983). When values pile up at a known limit, we can write a model for the outcome we would have seen without the limit and estimate it from the pile-up and the rest, and that is the Tobit model (Tobin, 1958)[45]. Both assume the mechanism is exactly what the model says, a known endpoint or a known limit, and neither is a general answer to confidentiality treatment: a swapped or noise-injected value is not a censored one, and a model built for limits has nothing to say about it.

The first exercise is... actually reading (yourself or your agents - helpful if it's in a language you don't speak), and it comes before any model: the agency documents its release rules, and the documentation says which mechanism affected which variable in which year (U.S. Census Bureau, 2023). For top-coded earnings, the workable repair is a tail model (a distribution fitted to the upper part of the data and used to fill in above the limit) and it earns its keep only when its family, its threshold and its fit are defended for the application at hand; multiple

[44] The confidentiality mechanisms deserve their own definitions because the words look interchangeable and they aren't, actually. Swapping exchanges values, like earnings, between records that look similar on other variables, so each record is real but some are wearing a neighbour's number. Noise addition perturbs values with random error whose distribution the agency knows and we usually don't. Suppression comes in two rounds: primary suppression blanks the sensitive cell, like the average earnings of the three trainees in a small county, and secondary suppression blanks other cells too since a blank cell inside published row and column totals could otherwise be recovered by subtraction. The umbrella term for all of it is statistical disclosure control, and agencies publish their general approach while keeping some parameters secret on purpose, which is a constraint our analysis has to live with rather than an oversight to complain about - we can still complain tho. Search terms: "statistical disclosure control", "cell suppression", "data swapping", "top-coding".

[45] Interval regression and Tobit get confused because both involve limits, and the difference decides the assumptions we get. On one hand, interval regression knows each observation's endpoints, so it needs a distributional assumption only to spread probability inside the brackets (kind of a mild request). On the other hand, the Tobit model assumes a full latent outcome (usually $\sim N$) and its estimates lean on such normality harder than we would expect, so a heavy-tailed or skewed truth can move the results. If the question is about conditional quantiles rather than a latent mean, censored quantile regression is the standard option (Powell, 1986), and it still needs the censoring rule and the quantile restriction to be right, while heavy censoring can leave some quantiles unidentified no matter what we do. There's also a third "animal": tail models for top-codes. Jenkins et al. (2011) fit a flexible distribution (the generalised beta of the second kind, GB2) to the upper range and impute above the limit, and their multiple-imputation machinery brings the tail uncertainty into the standard errors, which is what makes the exercise conditional-on-model in a way we can inspect. Implementations: `intreg` and `tobit` in Stata, `survreg` in R; search terms: "interval regression", "Tobit model", "top-coded earnings", "GB2 distribution".

imputation above the top-code, for example, produces answers that are conditional on the tail model chosen, and different defensible tails give different inequality estimates (Jenkins et al., 2011). You will meet more general recipes in the wild (e.g. fit a Pareto tail above any top-code, or recover suppressed cells from the published margins with optimisation) and this is the part where I should stop short of recommending them: the case-by-case defence is the method, and the generic version I can give would simply skip it.

In this entry, the mechanism decides. Treating a top-code as if it were censoring, or noise-injected values as if they were true ones, mixes two release rules into one analysis, and the output comes without a warning. The exercise that protects us costs nothing but time: match what we do to what the documentation says was done.

P17. The official outcome arrives too late

Our training subsidy is up for renewal and the state has to decide this quarter whether it worked, but the official earnings series for this quarter will only be published next year. The missing object is the current value at the decision date: the number exists in the world, in payrolls already paid, and the statistical system hasn't finished counting it. Meanwhile the timely data pile up (e.g. unemployment claims, job postings, payroll-processor records) and the exercise of estimating the official number from them before it's released is called nowcasting (Giannone et al., 2008).

The standard tools to solve this problem are bridge equations, factor models and mixed-frequency models[46], and they have one thing in common in order to make them work: the information set must be dated. A bridge equation links the quarterly target to monthly indicators through a regression, a factor model compresses many indicators into a few common movements and reads the target off them, and mixed-frequency models let monthly and quarterly data live in one equation without forcing them to the same calendar (Bańbura et al., 2013). The shared "discipline" concerns what the model is allowed to see: for any date we nowcast, only the data releases that existed on that date can enter. Later revisions and indicators published afterwards were not available to the decision we're trying to inform.

We need to be careful because that discipline is also where we can get the evaluation wrong. If we test our nowcasting model against history using today's data, we commit a crime in two types of hindsight: revised values of the indicators - which are cleaner than what was on the screen at the time - and series that hadn't been released yet on the dates we're pretending to nowcast. A fair historical test recreates what was known at each date by using real-time vintages, meaning the data exactly as they stood on each past day[47] (Croushore and Stark,

[46]The simplest one is a bridge equation: you aggregate the monthly indicators to quarterly frequency and regress the target on them, so the "bridge" crosses from what's published monthly to what's published quarterly. A dynamic factor model assumes many indicators move together because a few underlying forces drive them, extracts those forces, and nowcasts from the forces rather than the indicators, which is why it digests dozens of series without falling apart; the version in Giannone et al. (2008) also updates its nowcast release by release, so you can read how much each morning's data changed the estimate, a decomposition the literature calls the "news". Mixed-frequency regressions (search "MIDAS") skip the aggregation step and let a quarterly variable sit on the left with monthly regressors on the right. Many of these models are written in state-space form and updated with a Kalman filter, which is handy when releases arrive on different calendars and the dataset has a ragged edge (Bańbura et al., 2013). Search terms: "nowcasting", "bridge equation", "dynamic factor model", "nowcast news decomposition".

[47]A vintage is a snapshot of a data series as it stood on a given date, and vintage archives exist because researchers kept needing yesterday's version of the truth: the Philadelphia Fed's Real-Time Data Set, built from the work in Croushore and Stark (2001), and the St. Louis Fed's ALFRED archive are the standard sources for US series. An evaluation that uses them is called a pseudo-real-time evaluation: pseudo because we run it now, real-time because every number the model sees is dated to what existed then. The evaluation has to close two separate gaps, and closing only one still overstates the model's accuracy: the indicator values must be the unrevised ones that were on record at the time, and the list of series itself must match what had been published

2001). A model that looks sharp against final data and ordinary against real-time data isn't being treated unfairly by the second test; the second test is actually the true one.

Nowcasting also has a neighbour it gets confused with. A nowcast estimates a current value nobody has measured yet, while backcasting (see previous entries) builds a historically comparable series for periods already measured under other rules; a model can do one well and the other badly because the two jobs need different things from it. And forecast accuracy is a statement about prediction only: an indicator can predict earnings movements while measuring nothing we would interpret as the effect of training, so a model that nowcasts well tells us what the number will be, and says nothing on why.

P18. The study population differs from the target

Our pilot went well and the state now wants to roll the programme out to everyone. The missing object is the target-population effect: the effect in the population the decision is about, which is not the population we studied. The two can differ because effects differ across people (e.g. training may help younger workers more than older ones) and our six pilot sites had their own mix of ages, industries and cities, while the state has another. A variable like age here is called an effect moderator, meaning a characteristic that changes the size of the effect, and the pilot's mix of moderators is baked into its headline number.

We have two[48] ways to move from the study to the target - standardisation and weighting - and they have four requirements in common. Standardisation computes the effect within subgroups (e.g. for younger and older workers separately) and then re-averages those effects using the target population's subgroup shares instead of the study's (Cole and Stuart, 2010). Weighting goes the other way around: it reweights the study so it resembles the target, giving more weight to the kinds of people the study under-represents, with weights built from the odds of being in the study rather than the target; these are inverse odds of sampling weights[49] (Westreich et al., 2017). Both need the study effect itself to be right (internal validity), the variables to mean the same thing in both datasets, the effect to carry across given the shared variables (conditional effect exchangeability), and every kind of person in the target to have some counterpart in the study (positivity, the same requirement we met with the missing respondents). The last one is the hard limit: if the pilot enrolled nobody over sixty, no weighting scheme produces the over-sixty effect because there is nothing to reweight.

The moderators we didn't measure get a different exercise. If we suspect the pilot sites differed from the state in something unrecorded (how motivated the local caseworkers were, for example) we can vary that unmeasured moderator in a sensitivity analysis and report how far the target effect could move. The calculation of Andrews and Oster (2019)[50] belongs here: it uses how

at all on that date. Search terms: "real-time vintage data", "pseudo real-time evaluation".

[48] If we are being really honest, there's a third that combines them: the doubly robust procedures. These stay consistent when either the outcome model or the participation model is right, though only inside the transport assumptions we already signed up for (Degtiar and Rose, 2023).

[49] Let's think in numbers. Suppose training raises earnings by 800 for under-forties and by 200 for over-forties. Our pilot was 70% under-forty so its headline effect is $0.7 \times 800 + 0.3 \times 200 = 620$. The state is 40% under-forty so the standardised target effect is $0.4 \times 800 + 0.6 \times 200 = 440$. Same study, same subgroup effects, and a headline that drops by a third when the mix changes. This arithmetic we are doing is the entire point of the exercise, and it only needed the subgroup effects and the target's shares. Inverse odds weighting reaches the same destination by another road: model each study person's probability of being in the study rather than the target given their characteristics, weight by the inverse odds, and the reweighted study behaves like the target on everything the model included. How to choose? Go for standardisation when the moderators are few and the subgroup effects are estimable, and weighting when the characteristics are many. Search terms: "standardization", "transportability", "inverse odds of sampling weights", "generalizability of trial results".

[50] The Andrews-Oster logic is: if the people and places that joined the study differ from the target in ways we can see, and those visible differences relate to the effect, then the visible selection gives us a yardstick for how much the invisible selection could move the result, under the assumption that the invisible part behaves like the visible

the observed characteristics relate to participation to calibrate how much the unobserved ones could move the effect, under its own selection model and a small-selection approximation, and it is a stress test under those terms, not an automatic bound, correction, or transport estimator.

A study that sits inside its target - our pilot counties are part of the state - and a target that is a different population altogether - another state asking whether our results apply there - are different problems with different names: generalisability for the first, transport for the second, and a study should say which one it is doing. And the weights themselves invite one very specific question. Two weighting schemes can produce similar-looking columns of numbers while resting on different assumptions, one about how the study was sampled and one about how effects carry across populations, so the thing to ask of any weight is which assumption it encodes because the numbers look the same either way.

P19. Treatment timing is uncertain or anticipated

When we set up the subsidy example, we assumed nobody responded before the policy took effect. This entry is for when that assumption is the problem. The subsidy legally starts in January, but the bill passed in October, the local news covered it, and training providers began advertising in November, so by the time January arrives, some of the response has already happened. The missing object is the behavioural onset of treatment: the date behaviour started responding which is not required to equal the legal date, and which no dataset column announces.

When the institutional record leaves several defensible onset dates (the bill's passage, the news coverage, the legal start, etc), the exercise is to run the analysis under each coding and report the set. Each coding changes real things: who counts as treated from when (the cohorts), which periods serve as untreated comparison, and the estimand itself, since "the effect of the subsidy from January" and "the effect of the subsidy from October" are different questions. This exercise is a stress test tailored to our design, and there is no general correction behind it: not knowing when treatment began is not a statistical problem with a standard fix, but rather it is a piece of missing institutional knowledge, and the codings are how we show what it costs us.

Does all of this ring a bell? Anticipation has its own tools when the early response is real and we are not talking about a dating error. People and firms respond before formal adoption whenever they can see the policy coming and acting early pays, and a period we labelled "untreated" then already contains part of the effect[51] (Malani and Reif, 2015). The group-time estimators most of us already use can account for this to some extent: we declare an "anticipation horizon" (a stated number of periods before adoption in which effects are allowed to exist), and the

part (the selection model) and that selection into the study is not too strong (the small-selection approximation). The answer moves with both assumptions, which is why the output reads as a stress test: it answers "how bad could unmeasured selection be if it resembles the measured kind", and it doesn't answer "what is the corrected target effect". The distinction between generalisability and transport also decides the machinery: when the study is nested in the target, the target's covariate distribution is known from the start; when the target is a separate population, its data enter as a second dataset with all of P15's compatibility questions attached. Search terms: "Andrews Oster external validity", "selection on observables and unobservables", "transportability".

[51] Malani and Reif (2015) studied tort reform, where physicians responded while reforms were still working through legislatures, and their point cuts two ways. A "pre-trend" can be anticipation rather than a violation of comparability: the groups were comparable, and the treated one started responding early. And anticipation contaminates the baseline. If the pre-period already contains part of the effect, the before-after comparison subtracts that early response from the later one. When the anticipation runs in the same direction as the effect, the estimate usually comes out too small, but with other dynamics the bias can go either way (Malani and Reif, 2015). The two readings ask for different responses, which is why the pre-trend discussion from P08's diagnostics and this entry's dating problem are neighbours rather than the same issue. Search terms: "anticipation effects", "pre-trends as anticipation", "announcement effects".

estimator moves its baseline back accordingly[52] (Callaway and Sant'Anna, 2021). We "declare" the horizon and we try our best to defend it with institutional facts, such as when the bill passed or when coverage began, rather than estimated from the data since the data can't distinguish early effects from differential trends on their own. We should do that.

There's one more common thing we can do: we drop the periods right around adoption and compare "cleanly before" with "cleanly after", which is called a donut. A donut can be informative, and it has two costs we should account for: it discards the transition data, and it can change the target contrast because an effect measured between "well before" and "well after" does not say much about the adoption period itself. A donut is a disclosed design choice, and should not be considered a universal repair for ambiguous timing. Our duty, present throughout this entry, is the same one: date the treatment with institutional evidence, show the codings, and treat the calendar as part of the design rather than as a setting nobody chose.

P20. One unit's treatment affects another unit

Our subsidy trains thousands of workers, and trained workers compete for jobs with untrained ones, so an untrained worker in a treated state is not untouched by the programme: some of what happens to their wages happens because their competitors got trained. Across the border the story continues. Firms in the comparison state recruit workers from ours, so the "untreated" comparison may contain a "diluted dose" of the treatment. In every entry so far we assumed that each person's outcome depends only on their own treatment. In this entry, such assumption holds no more. The missing object is the exposure mapping, which is the rule that says who can affect whom, over what distance, and through which channel. The data don't record that rule, so the mapping has to come from an argument about how the market works.

The designs that handle interference all work by restricting it to somewhere we can see[53]. One restriction allows effects to travel within groups but not between them (within a local labour market but not across markets). This is called partial interference[54] (Hudgens and Halloran, 2008). A more general version states outright which other units count for my outcome and how (such as workers in my commuting zone, weighted by how much we compete), and this is an exposure mapping in the formal sense. Once we declare one, contrasts such as "treated versus untreated, holding neighbours' exposure fixed" become estimable under the design (Aronow and Samii, 2017). This mapping actually does a bit more than just enable estimation. It defines the estimand: the effect of the subsidy now means "the effect given who counts as exposed to whom, under this specific mapping".

[52] In the Callaway and Sant'Anna (2021) framework, effects are estimated per cohort and period, and the comparison baseline is the cohort's last untreated period. Declaring an anticipation horizon of one period tells the estimator that the last untreated period is one earlier than the adoption date, so the baseline moves back and the adoption-adjacent period gets treated as possibly affected. The declaration costs us data since the baseline now sits further from the outcomes it anchors, and that is the price of admitting the response started early. Implementations: the `did` package in R and `csdid` in Stata, both with an anticipation argument. Search terms: "group-time average treatment effects", "anticipation horizon", "Callaway Sant'Anna".

[53] In experiments we can build this variation on purpose with a two-stage design, which first randomises the treatment intensity across groups and then randomises treatment within each group (Hudgens and Halloran, 2008), and that assignment structure is what gives us separate direct and spillover comparisons.

[54] The partial-interference setting comes from vaccines, where the spillover is entirely the point: your vaccination protects me. Hudgens and Halloran (2008) showed that with interference confined to groups, four effects become definable and estimable. The direct effect compares treated with untreated people in the same group. The indirect effect compares untreated people in a heavily treated group with untreated people in an untreated group, and it is the formal version of herd immunity. The total effect combines the two, and the overall effect compares whole groups under different treatment intensities. The same four questions translate to our subsidy study: the indirect effect is what happens to untrained workers' wages when many colleagues get trained, which for a labour economist is the displacement question. Search terms: "partial interference", "direct and indirect effects", "spillover effects", and the no-interference assumption's formal name, "SUTVA".

A common check re-runs the analysis with different reaches (spillovers allowed within 10 km, then 20, then 50). This kind of radius sensitivity is often called a ring analysis. What we get from it is sensitivity, meaning how much the estimates move as the assumed reach changes. What we don't get is the true spillover radius. Each radius is a different mapping, each mapping may define a different estimand, so the rings compare answers to related questions instead of closing in on one answer (Sävje, 2024). And when the mapping is misspecified, we end up with two errors at once[55], each with its own name: a definitional error, because we estimated a contrast other than the one we intended, and an identification error, because even that contrast may be estimated with bias.

When the network or the geography is too weakly known to defend one mapping, the better report is the set of estimates across the defensible mappings, so the reader can see what the missing knowledge costs us. A single exposure model hides that cost. And the assumption we started this entry with should be written down in our studies like any other. No interference is an assumption like parallel trends or missing at random, it can be wrong in ways that change the answer, and it spent nineteen entries being assumed without anyone saying so - which is exactly how an assumption ends up treated as a harmless default.

P21. The available measure doesn't settle the concept

Suppose the state asks whether our training programme improved workers' economic security, and we answer them with earnings. Earnings is a real variable, well measured for once, but it is still not the concept: economic security has something to do with stability, benefits, and how easily a job is lost, and a higher wage on a precarious contract may not be more of it. The missing object here is a defensible operationalisation, meaning a way of turning the concept we care about into a variable we can measure, plus the argument that the two line up. Back in P02 the variable existed and was poorly measured. In this entry the measurement can be perfect and the question still remains because what's in doubt is whether we (well, others) actually measured the right thing.

The main exercise in this entry is to check the measure against what theory expects from the concept. If we have prior evidence that earnings really tracks economic security, it should move with the things security moves with (savings, reported financial stress, job tenure) and it should not move one-for-one with things that are clearly not security (like a temporary overtime spike). This is called construct validation (Cronbach and Meehl, 1955). A stronger version of it measures several concepts with several methods at once and checks the full pattern: different measures of the same concept should agree, and measures of different concepts should disagree enough to count as different. This is the multitrait-multimethod comparison[56] (Campbell and

[55] The two-error decomposition is the useful lesson of Sävje (2024): when the declared mapping is wrong, part of the damage is that our estimand changed, and part is ordinary bias, and the two need different repairs - a better definition for the first, a better design for the second. This is also why ring analyses can disagree without any of them being broken, since each ring answers its own question. When we can't defend a single mapping, the practical move is to report estimates under several and to say which institutional knowledge would narrow the set, e.g. commuting-flow data that shows how far workers travel. Search terms: "exposure mapping misspecification", "ring analysis spillovers", "interference causal inference".

[56] Our friends at the Psychology department met this problem first. Cronbach and Meehl (1955) had to validate tests for things like anxiety and intelligence, where no direct benchmark exists, so they proposed checking the measure against the relationships theory predicts (if the concept should rise with X and fall with Y, the measure should too). Campbell and Fiske (1959) turned this into a table. The idea is, we measure several concepts with several methods, then ask for two things: different methods measuring the same concept should agree (convergent validity), and the same method measuring different concepts should not agree too much (discriminant validity). The second check catches a failure we in econ rarely test for. Two supposedly different concepts can move together just because the same method produced both, e.g. everything self-reported moving together because the same person answered everything on the same afternoon. Search terms: "construct validity", "multitrait-multimethod matrix", "convergent and discriminant validity", "common method bias".

Fiske, 1959). Both exercises can show us where a measure behaves as the concept should. Neither can prove the indicator is the concept, because behaving as expected is evidence, and identity is not something evidence of this kind can establish.

Economics runs on examples of this gap. Patent counts are an indicator related to innovation, publication measures indicate research output or visibility, employment is one observable labour-market outcome, and nominal expenditure is a monetary input measure. These are illustrations of common practice, and none of them is a validated proxy relationship. Whether any of them measures the intended concept depends on the question and on domain evidence, which is an argument each study has to make for its own case rather than inherit from the literature.

When the validation exercise does not support the measure, we have two options, and the less obvious one is often better. We can a) hunt for a better measure of the same concept, or b) we can revise the concept, and ask a question the evidence can answer (not "did training improve economic security" but "did training raise earnings", stated as exactly that[57]) (Adcock and Collier, 2001; Bollen and Lennox, 1991). A narrower question answered with a variable we understand beats a grand question answered with a variable we can only dream about.

P22. Data are revised after first release

This is the shortest entry in the guide because we've already met its tools. The revision triangle appeared with the classification breaks, and the vintages and the pseudo-real-time evaluation appeared with the nowcasts. What's left is the rule that ties them all together. The missing object is the vintage appropriate to our question, and "appropriate" changes with the question. A real-time policy analysis should use the latest data available at the decision date while a historical measurement exercise can use first releases, later releases, or one fixed vintage, as long as it says which. The final vintage is not automatic ground truth for a real-time question because the decision-maker we're studying never actually saw it (Croushore and Stark, 2001).

The diagnostics here are the exercises we already know. Vintage panels and revision triangles show how the estimates changed as information arrived, and a pseudo-real-time evaluation reruns our whole procedure with each date's information set (International Monetary Fund, 2018). What these expose is revision risk, meaning how much our answer depends on which release we used. What they can't do is decide which vintage is the right one for every estimand, because that choice belongs to the question, and the question is ours to state.

The eight verbs

After 22 problems, you might have noticed a pattern. The exercises we walked through are variations on a small set of responses, and the easiest way to see this is to write each response as a verb. We end up with eight of them. A single study usually does several at once, so treat the verbs as a way of reading applied work rather than boxes to sort papers into, and when a procedure combines them, we pull it apart and name the verb for each piece.

Bound it. If the data cannot give us one answer, report the range of answers they can support. The width of that range tells us something too: the wider it is, the more we would have to

[57] Adcock and Collier (2001) split operationalisation into steps: background concept, systematised concept, indicators, scores. The split is what tells us where the trouble is. Sometimes the indicator is fine and the concept was never pinned down, and no new variable repairs an undefined concept. Bollen and Lennox (1991) add a distinction that changes what validation means. Reflective indicators are effects of the concept (test answers reflect ability), while formative indicators build it (income, education and occupation together form socioeconomic status). The agreement logic that works for reflective measures is the wrong test for formative ones, whose components can disagree because they measure different parts on purpose. Search terms: "systematized concept Adcock Collier", "reflective and formative indicators", "measurement model".

assume to get to a precise answer (Tamer, 2010).

Stress it. Change an assumption and see when the conclusion changes. The useful result is not that the estimate moved, but how far we had to push the assumption before it did (Lu and White, 2014).

Reconstruct it. Sometimes we can use other information to fill in what we cannot observe directly. The important question is then whether the link we used to fill the gap is credible. A completed dataset does not make the missing information observed.

Combine, or transport it. Sometimes the information exists, just somewhere else. We can bring together datasets or populations, but doing so requires a reason to believe that what connects them is comparable. Successfully merging two files is not evidence that this assumption holds.

Generate reference worlds. Sometimes what is missing is the comparison itself: what would our statistic look like under some alternative world? We can generate that comparison in different ways, but what we learn from it depends on how that alternative world was constructed.

Change the question. Sometimes the original question asks more than the evidence can answer. In that case, the best option may be to ask a narrower question and say clearly that the target has changed.

Estimate, or regularise it. Sometimes the problem is that there is too little information to estimate the model reliably. Regularisation can make the estimate more stable, but it does so by adding structure. That structure is part of the answer.

Report, or audit it. Sometimes the missing object is the analytical record itself. Better reporting, shared code and computational checks can tell us what was done and whether we can reproduce it (Christensen and Miguel, 2018; Olken, 2015). They cannot tell us whether the original research design was valid.

Writing things this way forces us to say what we actually learned. Different exercises produce different kinds of evidence, and they should not all end up in a table labelled "robustness". They also come with a trade-off: if the data cannot answer the original question on their own, we either make an additional assumption, accept a less precise answer, or ask a different question.

This is where robustness checks can become ritualistic. A useful check starts with a specific problem and tells us something about it. A ritual check starts with a familiar method and works backwards. Some of this is probably driven by publication incentives and convention, but we should be careful about blaming referees for all of it.

There is good evidence that reasonable analytical choices can produce different results and that selective reporting affects what gets published (Steegen et al., 2016; Simonsohn et al., 2020). We also know something about what editors and referees value and about selection during publication (Berk et al., 2017; Brodeur et al., 2023). What we do not know is how much of the robustness machinery in published papers is there because referees asked for it. That story is plausible, but the evidence does not get us that far.

The same point applies to reproducibility. Being able to rerun an analysis is quite useful, but it does not settle whether the design was good in the first place or whether the finding would survive new data (Chang and Li, 2022). Reproducibility can close one gap while leaving the important one untouched.

So when something we need is missing, start there. Say what is missing and work out what the available evidence can still tell us. Sometimes the answer will be weaker or narrower than the one we wanted. That is fine. The standard robustness table can wait.

In the end, it is quite a simple rule: say what the exercise does, say what you had to assume to do it, and say what is still missing. A robustness check is useful when it changes what we know, not when it merely gives us another column for the table.

Any comments, suggestions or corrections, please send me an email at b.gietner@gmail.com. Fable-5 helped with the literature review and gpt-5.6-sol audited the text (twice just to be sure).

References

Alberto Abadie. Using Synthetic Controls: Feasibility, Data Requirements, and Methodological Aspects. *Journal of Economic Literature*, 59(2):391–425, 2021. doi:10.1257/jel.20191450.

Alberto Abadie, Alexis Diamond, and Jens Hainmueller. Synthetic Control Methods for Comparative Case Studies: Estimating the Effect of California's Tobacco Control Program. *Journal of the American Statistical Association*, 105(490):493–505, 2010. doi:10.1198/jasa.2009.ap08746.

Robert Adcock and David Collier. Measurement Validity: A Shared Standard for Qualitative and Quantitative Research. *American Political Science Review*, 95(3):529–546, 2001. doi:10.1017/S0003055401003100.

Dennis J. Aigner. Regression with a Binary Independent Variable Subject to Errors of Observation. *Journal of Econometrics*, 1(1):49–59, 1973. doi:10.1016/0304-4076(73)90005-5.

Paul S. Albert and Lori E. Dodd. A Cautionary Note on the Robustness of Latent Class Models for Estimating Diagnostic Error without a Gold Standard. *Biometrics*, 60(2):427–435, 2004. doi:10.1111/j.0006-341X.2004.00187.x.

Joseph G. Altonji, Todd E. Elder, and Christopher R. Taber. Selection on Observed and Unobserved Variables: Assessing the Effectiveness of Catholic Schools. *Journal of Political Economy*, 113(1):151–184, 2005. doi:10.1086/426036.

Isaiah Andrews and Emily Oster. A Simple Approximation for Evaluating External Validity Bias. *Economics Letters*, 178:58–62, 2019. doi:10.1016/j.econlet.2019.02.020.

Isaiah Andrews, James H. Stock, and Liyang Sun. Weak Instruments in Instrumental Variables Regression: Theory and Practice. *Annual Review of Economics*, 11:727–753, 2019. doi:10.1146/annurev-economics-080218-025643.

Joshua D. Angrist and Jörn-Steffen Pischke. The Credibility Revolution in Empirical Economics: How Better Research Design Is Taking the Con out of Econometrics. *Journal of Economic Perspectives*, 24(2):3–30, 2010. doi:10.1257/jep.24.2.3.

Dmitry Arkhangelsky, Susan Athey, David A. Hirshberg, Guido W. Imbens, and Stefan Wager. Synthetic Difference-in-Differences. *American Economic Review*, 111(12):4088–4118, 2021. doi:10.1257/aer.20190159.

Peter M. Aronow and Cyrus Samii. Estimating Average Causal Effects Under General Interference, with Application to a Social Network Experiment. *The Annals of Applied Statistics*, 11 (4):1912–1947, 2017. doi:10.1214/16-AOAS1005.

Susan Athey and Guido W. Imbens. The Econometrics of Randomized Experiments. In Abhijit Vinayak Banerjee and Esther Duflo, editors, *Handbook of Economic Field Experiments*, volume 1, pages 73–140. Elsevier, 2017. doi:10.1016/bs.hefe.2016.10.003.

Marta Bańbura, Domenico Giannone, Michele Modugno, and Lucrezia Reichlin. Now-Casting and the Real-Time Data Flow. In Graham Elliott, Clive W. J. Granger, and Allan Timmermann, editors, *Handbook of Economic Forecasting*, volume 2, pages 195–237. Elsevier, 2013. doi:10.1016/B978-0-444-53683-9.00004-9.

Heejung Bang and James M. Robins. Doubly Robust Estimation in Missing Data and Causal Inference Models. *Biometrics*, 61(4):962–973, 2005. doi:10.1111/j.1541-0420.2005.00377.x.

Jonathan W. Bartlett, Shaun R. Seaman, Ian R. White, and James R. Carpenter. Multiple Imputation of Covariates by Fully Conditional Specification: Accommodating the Substantive Model. *Statistical Methods in Medical Research*, 24(4):462–487, 2015. doi:10.1177/0962280214521348.

Eli Ben-Michael, Avi Feller, and Jesse Rothstein. The Augmented Synthetic Control

Method. *Journal of the American Statistical Association*, 116(536):1789–1803, 2021. doi:10.1080/01621459.2021.1929245.

Jonathan B. Berk, Campbell R. Harvey, and David Hirshleifer. How to Write an Effective Referee Report and Improve the Scientific Review Process. *Journal of Economic Perspectives*, 31(1):231–244, 2017. doi:10.1257/jep.31.1.231.

Kenneth A. Bollen and Richard Lennox. Conventional Wisdom on Measurement: A Structural Equation Perspective. *Psychological Bulletin*, 110(2):305–314, 1991. doi:10.1037/0033-2909.110.2.305.

Christopher R. Bollinger. Bounding Mean Regressions When a Binary Regressor Is Mismeasured. *Journal of Econometrics*, 73(2):387–399, 1996. doi:10.1016/S0304-4076(95)01730-5.

John Bound and Alan B. Krueger. The Extent of Measurement Error in Longitudinal Earnings Data: Do Two Wrongs Make a Right? *Journal of Labor Economics*, 9(1):1–24, 1991. doi:10.1086/298256.

John Bound, Charles Brown, Greg J. Duncan, and Willard L. Rodgers. Evidence on the Validity of Cross-Sectional and Longitudinal Labor Market Data. *Journal of Labor Economics*, 12(3): 345–368, 1994. URL `https://ideas.repec.org/a/ucp/jlabec/v12y1994i3p345-68.html`.

John Bound, Charles Brown, and Nancy Mathiowetz. Measurement Error in Survey Data. In James J. Heckman and Edward E. Leamer, editors, *Handbook of Econometrics*, volume 5, chapter 59, pages 3705–3843. Elsevier, 2001. doi:10.1016/S1573-4412(01)05012-7.

Abel Brodeur, Scott Carrell, David Figlio, and Lester Lusher. Unpacking P-hacking and Publication Bias. *American Economic Review*, 113(11):2974–3002, 2023. doi:10.1257/aer.20210795.

Brantly Callaway and Pedro H. C. Sant'Anna. Difference-in-Differences with Multiple Time Periods. *Journal of Econometrics*, 225(2):200–230, 2021. doi:10.1016/j.jeconom.2020.12.001.

A. Colin Cameron and Douglas L. Miller. A Practitioner's Guide to Cluster-Robust Inference. *Journal of Human Resources*, 50(2):317–372, 2015. doi:10.3368/jhr.50.2.317.

A. Colin Cameron, Jonah B. Gelbach, and Douglas L. Miller. Bootstrap-Based Improvements for Inference with Clustered Errors. *The Review of Economics and Statistics*, 90(3):414–427, 2008. doi:10.1162/rest.90.3.414.

Donald T. Campbell and Donald W. Fiske. Convergent and Discriminant Validation by the Multitrait-Multimethod Matrix. *Psychological Bulletin*, 56(2):81–105, 1959. doi:10.1037/h0046016.

Raymond J. Carroll, David Ruppert, Leonard A. Stefanski, and Ciprian M. Crainiceanu. *Measurement Error in Nonlinear Models: A Modern Perspective*. Chapman & Hall/CRC, Boca Raton, FL, 2 edition, 2006. ISBN 9781584886334. doi:10.1201/9781420010138.

Karim Chalak. Identification of Average Effects under Magnitude and Sign Restrictions on Confounding. *Quantitative Economics*, 10(4):1619–1657, 2019. doi:10.3982/QE689.

Andrew C. Chang and Phillip Li. Is Economics Research Replicable? Sixty Published Papers From Thirteen Journals Say "Often Not". *Critical Finance Review*, 11(1):185–206, 2022. doi:10.1561/104.00000053.

Wendy K. Tam Cho and Charles F. Manski. Cross-Level/Ecological Inference. In *The Oxford Handbook of Political Methodology*. Oxford University Press, 2008. ISBN 9780199286546. doi:10.1093/oxfordhb/9780199286546.003.0024.

Garret Christensen and Edward Miguel. Transparency, Reproducibility, and the Credibility of Economics Research. *Journal of Economic Literature*, 56(3):920–980, 2018. doi:10.1257/jel.20171350.

EunYi Chung and Joseph P. Romano. Exact and Asymptotically Robust Permutation Tests.

The Annals of Statistics, 41(2):484–507, 2013. doi:10.1214/13-AOS1090.

Carlos Cinelli and Chad Hazlett. Making Sense of Sensitivity: Extending Omitted Variable Bias. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 82(1): 39–67, 2020. doi:10.1111/rssb.12348.

Stephen R. Cole and Elizabeth A. Stuart. Generalizing Evidence From Randomized Clinical Trials to Target Populations: The ACTG 320 Trial. *American Journal of Epidemiology*, 172 (1):107–115, 2010. doi:10.1093/aje/kwq084.

J. R. Cook and Leonard A. Stefanski. Simulation-Extrapolation Estimation in Parametric Measurement Error Models. *Journal of the American Statistical Association*, 89(428):1314–1328, 1994. doi:10.1080/01621459.1994.10476871.

Lee J. Cronbach and Paul E. Meehl. Construct Validity in Psychological Tests. *Psychological Bulletin*, 52(4):281–302, 1955. doi:10.1037/h0040957.

Dean Croushore and Tom Stark. A Real-Time Data Set for Macroeconomists. *Journal of Econometrics*, 105(1):111–130, 2001. doi:10.1016/S0304-4076(01)00072-0.

Richard K. Crump, V. Joseph Hotz, Guido W. Imbens, and Oscar A. Mitnik. Dealing with Limited Overlap in Estimation of Average Treatment Effects. *Biometrika*, 96(1):187–199, 2009. doi:10.1093/biomet/asn055.

Jan de Haan and Heymerik A. van der Grient. Eliminating Chain Drift in Price Indexes Based on Scanner Data. *Journal of Econometrics*, 161(1):36–46, 2011. doi:10.1016/j.jeconom.2010.09.004.

Angus Deaton. Quality, Quantity, and Spatial Variation of Price. *American Economic Review*, 78(3):418–430, 1988. URL `https://www.jstor.org/stable/1809142`. Stable JSTOR identifier 1811174.

Irina Degtiar and Sherri Rose. A Review of Generalizability and Transportability. *Annual Review of Statistics and Its Application*, 10:501–524, 2023. doi:10.1146/annurev-statistics-042522-103837.

William G. Dewald, Jerry G. Thursby, and Richard G. Anderson. Replication in Empirical Economics: The Journal of Money, Credit and Banking Project. *American Economic Review*, 76(4):587–603, 1986. URL `https://www.jstor.org/stable/1806061`. Stable JSTOR identifier 1806061.

Peng Ding and Fan Li. Causal Inference: A Missing Data Perspective. *Statistical Science*, 33 (2):214–237, 2018. doi:10.1214/18-STS645.

Jessie K. Edwards, Stephen R. Cole, and Daniel Westreich. All Your Data Are Always Missing: Incorporating Bias Due to Measurement Error into the Potential Outcomes Framework. *International Journal of Epidemiology*, 44(4):1452–1459, 2015. doi:10.1093/ije/dyu272.

Ivan P. Fellegi and Alan B. Sunter. A Theory for Record Linkage. *Journal of the American Statistical Association*, 64(328):1183–1210, 1969. doi:10.1080/01621459.1969.10501049.

David Firth. Bias Reduction of Maximum Likelihood Estimates. *Biometrika*, 80(1):27–38, 1993. doi:10.1093/biomet/80.1.27.

Matthew P. Fox, Timothy L. Lash, and Sander Greenland. A Method to Automate Probabilistic Sensitivity Analyses of Misclassified Binary Variables. *International Journal of Epidemiology*, 34(6):1370–1376, 2005. doi:10.1093/ije/dyi184.

Constantine E. Frangakis and Donald B. Rubin. Principal Stratification in Causal Inference. *Biometrics*, 58(1):21–29, 2002. doi:10.1111/j.0006-341X.2002.00021.x.

Andrew Gelman. Prior Distributions for Variance Parameters in Hierarchical Models. *Bayesian Analysis*, 1(3):515–533, 2006. doi:10.1214/06-BA117A.

Domenico Giannone, Lucrezia Reichlin, and David Small. Nowcasting: The Real-Time Informational Content of Macroeconomic Data. *Journal of Monetary Economics*, 55(4):665–676, 2008. doi:10.1016/j.jmoneco.2008.05.010.

Sander Greenland. Ecologic versus Individual-Level Sources of Bias in Ecologic Estimates of Contextual Health Effects. *International Journal of Epidemiology*, 30(6):1343–1350, 2001. doi:10.1093/ije/30.6.1343.

Zvi Griliches and Jerry A. Hausman. Errors in Variables in Panel Data. *Journal of Econometrics*, 31(1):93–118, 1986. doi:10.1016/0304-4076(86)90058-8.

Jerry Hausman. Mismeasured Variables in Econometric Analysis: Problems from the Right and Problems from the Left. *Journal of Economic Perspectives*, 15(4):57–67, 2001. doi:10.1257/jep.15.4.57.

Jerry A. Hausman, Jason Abrevaya, and F. M. Scott-Morton. Misclassification of the Dependent Variable in a Discrete-Response Setting. *Journal of Econometrics*, 87(2):239–269, 1998. doi:10.1016/S0304-4076(98)00015-3.

James J. Heckman. Sample Selection Bias as a Specification Error. *Econometrica*, 47(1):153–161, 1979. doi:10.2307/1912352.

Georg Heinze and Michael Schemper. A Solution to the Problem of Separation in Logistic Regression. *Statistics in Medicine*, 21(16):2409–2419, 2002. doi:10.1002/sim.1047.

Joel L. Horowitz and Charles F. Manski. Censoring of Outcomes and Regressors due to Survey Nonresponse: Identification and Estimation Using Weights and Imputations. *Journal of Econometrics*, 84(1):37–58, 1998. doi:10.1016/S0304-4076(97)00077-8.

Chanelle J. Howe, Lauren E. Cain, and Joseph W. Hogan. Are All Biases Missing Data Problems? *Current Epidemiology Reports*, 2(3):162–171, 2015. doi:10.1007/s40471-015-0050-8.

Michael G. Hudgens and M. Elizabeth Halloran. Toward Causal Inference With Interference. *Journal of the American Statistical Association*, 103(482):832–842, 2008. doi:10.1198/016214508000000292.

Siu L. Hui and Stephen D. Walter. Estimating the Error Rates of Diagnostic Tests. *Biometrics*, 36(1):167–171, 1980. doi:10.2307/2530508.

Guido W. Imbens. Sensitivity to Exogeneity Assumptions in Program Evaluation. *American Economic Review*, 93(2):126–132, 2003. doi:10.1257/000282803321946921.

Atsushi Inoue and Gary Solon. Two-Sample Instrumental Variables Estimators. *The Review of Economics and Statistics*, 92(3):557–561, 2010. doi:10.1162/REST_a_00011.

International Monetary Fund. *Quarterly National Accounts Manual: 2017 Edition*. International Monetary Fund, Washington, DC, 2018. ISBN 9781475589870. doi:10.5089/9781475589870.069.

Intersecretariat Working Group on Price Statistics. *Consumer Price Index Manual: Concepts and Methods*. International Monetary Fund, Washington, DC, 2020. ISBN 9781484354841. doi:10.5089/9781484354841.069.

Stephen P. Jenkins, Richard V. Burkhauser, Shuaizhang Feng, and Jeff Larrimore. Measuring Inequality Using Censored Data: A Multiple-Imputation Approach to Estimation and Inference. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 174(1):63–81, 2011. doi:10.1111/j.1467-985X.2010.00655.x.

Gary King. *A Solution to the Ecological Inference Problem: Reconstructing Individual Behavior from Aggregate Data*. Princeton University Press, Princeton, 1997. doi:10.1515/9781400849208.

Gary King and Langche Zeng. Logistic Regression in Rare Events Data. *Political Analysis*, 9

(2):137–163, 2001. doi:10.1093/oxfordjournals.pan.a004868.

Brian Krauth. Bounding a Linear Causal Effect Using Relative Correlation Restrictions. *Journal of Econometric Methods*, 5(1):117–141, 2016. doi:10.1515/jem-2013-0013.

P. Lahiri and Michael D. Larsen. Regression Analysis With Linked Data. *Journal of the American Statistical Association*, 100(469):222–230, 2005. doi:10.1198/016214504000001277.

David S. Lee. Training, Wages, and Sample Selection: Estimating Sharp Bounds on Treatment Effects. *The Review of Economic Studies*, 76(3):1071–1102, 2009. doi:10.1111/j.1467-937X.2009.00536.x.

Fan Li, Kari Lock Morgan, and Alan M. Zaslavsky. Balancing Covariates via Propensity Score Weighting. *Journal of the American Statistical Association*, 113(521):390–400, 2018. doi:10.1080/01621459.2016.1260466.

Marc Lipsitch, Eric Tchetgen Tchetgen, and Ted Cohen. Negative Controls: A Tool for Detecting Confounding and Bias in Observational Studies. *Epidemiology*, 21(3):383–388, 2010. doi:10.1097/EDE.0b013e3181d61eeb.

Roderick J. A. Little. Pattern-Mixture Models for Multivariate Incomplete Data. *Journal of the American Statistical Association*, 88(421):125–134, 1993. doi:10.1080/01621459.1993.10594302.

Roderick J. A. Little and Donald B. Rubin. *Statistical Analysis with Missing Data*. Wiley, 3 edition, 2019. doi:10.1002/9781119482260.

Xun Lu and Halbert White. Robustness Checks and Robustness Tests in Applied Economics. *Journal of Econometrics*, 178(1):194–206, 2014. doi:10.1016/j.jeconom.2013.08.016.

Anup Malani and Julian Reif. Interpreting Pre-trends as Anticipation: Impact on Estimated Treatment Effects from Tort Reform. *Journal of Public Economics*, 124:1–17, 2015. doi:10.1016/j.jpubeco.2015.01.001.

Xiao-Li Meng. Multiple-Imputation Inferences with Uncongenial Sources of Input. *Statistical Science*, 9(4):538–558, 1994. doi:10.1214/ss/1177010269.

Wang Miao, Zhi Geng, and Eric J. Tchetgen Tchetgen. Identifying Causal Effects with Proxy Variables of an Unmeasured Confounder. *Biometrika*, 105(4):987–993, 2018. doi:10.1093/biomet/asy038.

Geert Molenberghs, Caroline Beunckens, Cristina Sotto, and Michael G. Kenward. Every Missingness Not at Random Model Has a Missingness at Random Counterpart with Equal Fit. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70(2):371–388, 2008. doi:10.1111/j.1467-9868.2007.00640.x.

Benjamin A. Olken. Promises and Perils of Pre-analysis Plans. *Journal of Economic Perspectives*, 29(3):61–80, 2015. doi:10.1257/jep.29.3.61.

Emily Oster. Unobservable Selection and Coefficient Stability: Theory and Evidence. *Journal of Business & Economic Statistics*, 37(2):187–204, 2019. doi:10.1080/07350015.2016.1227711.

Zhuan Pei, Jörn-Steffen Pischke, and Hannes Schwandt. Poorly Measured Confounders Are More Useful on the Left than on the Right. *Journal of Business & Economic Statistics*, 37 (2):205–216, 2019. doi:10.1080/07350015.2018.1462710.

Justin R. Pierce and Peter K. Schott. A Concordance between Ten-Digit U.S. Harmonized System Codes and SIC/NAICS Product Classes and Industries. *Journal of Economic and Social Measurement*, 37(1-2):61–96, 2012. doi:10.3233/JEM-2012-0351.

James L. Powell. Censored Regression Quantiles. *Journal of Econometrics*, 32(1):143–155, 1986. doi:10.1016/0304-4076(86)90016-3.

Ashesh Rambachan and Jonathan Roth. A More Credible Approach to Parallel Trends. *The*

Review of Economic Studies, 90(5):2555–2591, 2023. doi:10.1093/restud/rdad018.

Susanne Rässler. Data Fusion: Identification Problems, Validity, and Multiple Imputation. *Austrian Journal of Statistics*, 33(1&2):153–171, 2004. doi:10.17713/ajs.v33i1&2.436.

Geert Ridder and Robert Moffitt. The Econometrics of Data Combination. In *Handbook of Econometrics*, volume 6B, chapter 75, pages 5469–5547. Elsevier, 2007. doi:10.1016/S1573-4412(07)06075-8.

James M. Robins, Andrea Rotnitzky, and Lue Ping Zhao. Estimation of Regression Coefficients When Some Regressors Are Not Always Observed. *Journal of the American Statistical Association*, 89(427):846–866, 1994. doi:10.1080/01621459.1994.10476818.

W. S. Robinson. Ecological Correlations and the Behavior of Individuals. *American Sociological Review*, 15(3):351–357, 1950. doi:10.2307/2087176.

Sherwin Rosen. Hedonic Prices and Implicit Markets: Product Differentiation in Pure Competition. *Journal of Political Economy*, 82(1):34–55, 1974. doi:10.1086/260169.

Paul R. Rosenbaum. Sensitivity Analysis for Certain Permutation Inferences in Matched Observational Studies. *Biometrika*, 74(1):13–26, 1987. doi:10.1093/biomet/74.1.13.

Paul R. Rosenbaum. *Observational Studies*. Springer Series in Statistics. Springer, New York, 2 edition, 2002. ISBN 978-0-387-98967-9. doi:10.1007/978-1-4757-3692-2.

Jonathan Roth. Pretest with Caution: Event-Study Estimates after Testing for Parallel Trends. *American Economic Review: Insights*, 4(3):305–322, 2022. doi:10.1257/aeri.20210236.

Donald B. Rubin. Inference and Missing Data. *Biometrika*, 63(3):581–592, 1976. doi:10.1093/biomet/63.3.581.

Donald B. Rubin. *Multiple Imputation for Nonresponse in Surveys*. John Wiley & Sons, New York, 1987. doi:10.1002/9780470316696.

Mauricio Sadinle. Bayesian Estimation of Bipartite Matchings for Record Linkage. *Journal of the American Statistical Association*, 112(518):600–612, 2017. doi:10.1080/01621459.2016.1148612.

Fredrik Sävje. Causal Inference with Misspecified Exposure Mappings: Separating Definitions and Assumptions. *Biometrika*, 111(1):1–15, 2024. doi:10.1093/biomet/asad019.

Daniel O. Scharfstein, Andrea Rotnitzky, and James M. Robins. Adjusting for Nonignorable Drop-Out Using Semiparametric Nonresponse Models. *Journal of the American Statistical Association*, 94(448):1096–1120, 1999. doi:10.1080/01621459.1999.10473862.

Uri Simonsohn, Joseph P. Simmons, and Leif D. Nelson. Specification Curve Analysis. *Nature Human Behaviour*, 4(11):1208–1214, 2020. doi:10.1038/s41562-020-0912-z.

Douglas Staiger and James H. Stock. Instrumental Variables Regression with Weak Instruments. *Econometrica*, 65(3):557–586, 1997. doi:10.2307/2171753.

Sara Steegen, Francis Tuerlinckx, Andrew Gelman, and Wolf Vanpaemel. Increasing Transparency Through a Multiverse Analysis. *Perspectives on Psychological Science*, 11(5):702–712, 2016. doi:10.1177/1745691616658637.

Mark B. Stewart. On Least Squares Estimation When the Dependent Variable Is Grouped. *The Review of Economic Studies*, 50(4):737–753, 1983. doi:10.2307/2297773.

Elie Tamer. Partial Identification in Econometrics. *Annual Review of Economics*, 2:167–195, 2010. doi:10.1146/annurev.economics.050708.143401.

James Tobin. Estimation of Relationships for Limited Dependent Variables. *Econometrica*, 26 (1):24–36, 1958. doi:10.2307/1907382.

U.S. Census Bureau. Current Population Survey, 2023 Annual Social and Economic (ASEC)

Supplement: Technical Documentation. Technical report, U.S. Census Bureau, 2023. URL https://www2.census.gov/programs-surveys/cps/techdocs/cpsmar23.pdf. See section 8, Data Disclosure Avoidance Techniques.

Tyler J. VanderWeele and Peng Ding. Sensitivity Analysis in Observational Research: Introducing the E-Value. *Annals of Internal Medicine*, 167(4):268–274, 2017. doi:10.7326/M16-2607.

Daniel Westreich, Jessie K. Edwards, Catherine R. Lesko, Elizabeth A. Stuart, and Stephen R. Cole. Transportability of Trial Results Using Inverse Odds of Sampling Weights. *American Journal of Epidemiology*, 186(8):1010–1014, 2017. doi:10.1093/aje/kwx164.

Qingyuan Zhao, Jingshu Wang, Wes Spiller, Jack Bowden, and Dylan S. Small. Two-Sample Instrumental Variable Analyses Using Heterogeneous Samples. *Statistical Science*, 34(2): 317–333, 2019. doi:10.1214/18-STS692.