Chapter 5

Bridging Cases

The previous chapter (Chapter 4) introduced a suite of tools for modeling differential item functioning (DIF) between survey respondents. These tools can help researchers effectively use the information that respondents provide to produce accurate measurements of relevant concepts or quantities of interest. But respondent differences are only one challenge to the goal of deploying effective surveys to measure concepts across cases. *Respondents—who themselves vary in how they respond to survey questions—also provide responses that are specific to the contexts in which they operate.*

For instance, a public opinion survey of presidential approval in Russia provides fundamentally different information than a similar survey in the United States because the factors that affect presidential approval will likely vary substantially across these two disparate cases. Even within a single country, the meaning of such a survey question could vary considerably across time. For example, we are currently in a period of low public approval of US presidents, relative to the past. While this distinction could reflect what we think it does—that the US public was happier with earlier presidents than they are with our current options—it is also possible that both the people who respond to these surveys, and the context in which they operate, changed. Indeed, in some cases the respondent pool has turned over entirely: the people who answered presidential approval polls in the early 1960s are, for the most part, no longer alive. These changes make approval ratings for, say, John Kennedy and Joe Biden, largely incomparable.

Like the problems we address in the previous chapter, this is an inter-respondent comparability issue.[1] But *inter-respondent incomparability becomes a "bridging" problem when discrete sets of respondents provide information about different cases in systematically different ways.*

This facet of cross-case surveys makes it especially difficult to ensure that one can interpret responses consistently across respondents and, in turn, across cases. In general, the tools that we introduced in the previous chapter can help us to make respondents comparable, but our measurements will still lack comparability if "un-bridged" groups of respondents assess different cases. In such situations, we can use our toolkit to make respondents lo-

[1]These differences can also be driven by case-level variables. Imagine, for instance, that expressing negative views of public figures was less socially permissible in the past, causing individual respondents to systematically assess such figures 10 points higher in the 1960s than they do now, everything else being equal.

cally comparable, but we need respondents who assess multiple cases—who *bridge* the case space—to generally adjust for inter-respondent differences and achieve global comparability. For the most part, we field cross-case surveys specifically to compare cases. But we cannot convincingly compare cases without effective bridging. Deploying cross-case surveys without carefully addressing bridging is, therefore, at best, a waste of time, and at worst harmfully misleading!

Statement of the Problem
We typically conduct cross-case surveys to compare cases, but the respondents that evaluate one case often answer questions differently from those that evaluate another, exhibiting case-driven DIF and making their responses fundamentally non-comparable.

To further illustrate the bridging problem, consider a peculiar hypothetical survey: an essay-based exam given to high-school students around the world. Imagine that this exam is identical everywhere and perfectly translated to each relevant language, but that local language arts teachers assess students' responses in each country. These teachers do not receive exam-specific training and bring their local educational expectations to the grading process. Multiple (say three) teachers in each country evaluate each exam, and exams are randomly distributed so that there is a lot of overlap between graders in each country. Each grader ends up co-grading a handful of papers with every other teacher for that country.

Of course, within each country, some graders will be "hard" graders, while others will be—relatively speaking—pushovers. The item response theory (IRT) framework that we introduced in the previous chapter has a way of dealing with how strictly each individual teacher grades. As long as multiple teachers grade each essay, IRT can use the pattern of grades that teachers assign to estimate *threshold parameters* for each teacher. It can learn, in other words, that a given teacher has a higher threshold for assigning an "A" to an exam than another and can distinguish the "hard" graders from the "easy" ones. In turn, we can normalize exam scores based on the set of teachers who graded each given exam, using threshold estimates to adjust grades so that students who pulled difficult evaluators obtain higher adjustments to their scores than the lucky students who received easier graders. A larger adjustment does not guarantee a higher final score, however—a weaker student who drew a harsh grader should still finish below a stronger student who drew an easy one.

Yet, given the way that teachers cluster into countries, we will never observe exams that were graded by teachers hailing from multiple different countries. If country-level factors—educational norms, culture, and resources—affect how exacting teachers are on average, then the distribution of hard and easy graders will vary systematically across countries, but the IRT model will lack the information to incorporate these cross-country differences into threshold estimates, and will have to rely on the assumption that a hard grader in France is about as tough as a hard grader in Japan, and so on. When this assumption is wrong, as it likely is in this example, then cross-country differences in grades—even after we adjust them based on estimated within-country grader thresholds—will reflect different grading scales and will be fundamentally incomparable.

While the bridge problem can affect all sorts of survey questions and related tasks, it is often especially acute for the general class of problems that are the focus of this book: the cross-case measurement of latent constructs. For instance, consider the V-Dem project

or any other cross-case expert survey. V-Dem respondents are, for the most part, experts on specific cases. Indeed, V-Dem's slogan is "Global standards, local knowledge." Thus, certain sets of respondents provide survey responses about different cases. In the extreme, each country is evaluated by a discrete set of respondents. In the V-Dem context, where each expert assesses a long time-period, applying the Chapter 4 toolkit to *un-bridged* sets of local experts would allow us to say a lot about democratic institutions within countries over time. But it would not allow us to say very much about differences across countries. We would be left with half of V-Dem's slogan—effectively summarizing local knowledge, but with no hope of obtaining global standards. Similar problems arise when one tries to construct what are called ideological common spaces (consistent measures of ideology across political bodies and actors) and on ratings sites (where geographically, temporally, or otherwise disparate consumers provide ratings to different products or services). Mass surveys often focus on summarizing particular populations—for instance, the attitude of the typical voter to a sitting president. But whenever a researcher wants to compare such measures across cases, the bridge problem rears its head.

Solution Punchline Points

1. Bridge observations are assessments of cases provided by respondents who, otherwise, assess disparate sets of cases.
2. IRT models can use information from bridge observations to adjust for DIF and produce measures that have consistent meanings across cases, but they will generate misleading measures when bridges are lacking.
3. Bridge observations are often difficult to collect, so efficiently selecting bridges can save researchers both time and money.
4. Cross-case survey researchers and downstream data consumers should carefully check the bridge density of their data, assess cross-case face validity, and use tools like posterior predictive checks to evaluate whether their measures suffer from insufficient bridging.

This chapter examines the bridging problem in detail. First, we use a toy example to demonstrate how ignoring the bridge problem can lead one astray. We then show how one can add bridge observations and apply measurement modeling techniques to achieve cross-case consistency in measurement. We extend this discussion to the real world and take a tour of the bridging problem in different survey contexts. Using applied examples from expert and mass surveys, ideology estimation, and customer-service ratings, we demonstrate real-life bridging issues, discuss practical strategies to bridge different sorts of survey data, and further demonstrate how bringing bridges—and measurement modeling techniques—to bear can improve cross-case measurement. We next develop rules of thumb and automated techniques for achieving sufficient bridging at minimal cost by strategically selecting highly informative bridges. Finally, we discuss general techniques for assessing data for bridging issues and consider the problem of whether a particular survey contains sufficient cross-case

bridging. Chapter 6 then extends this discussion by conducting a deep dive into one, generally applicable, bridging strategy: asking respondents to answer questions about "anchoring vignettes."

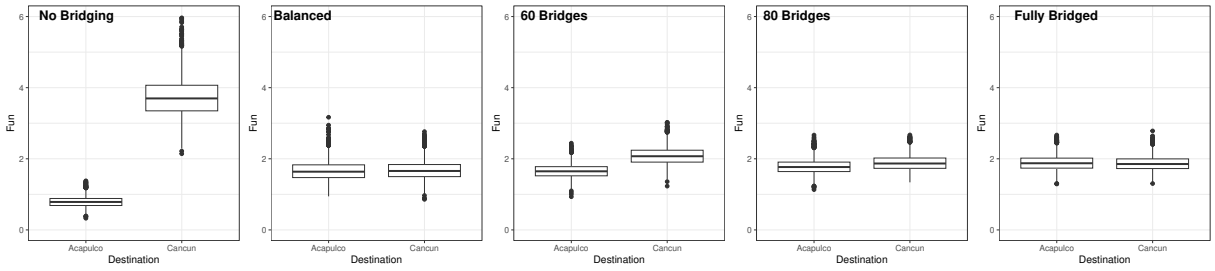
5.1 What Happens When We Lack Bridges?

Imagine you are deciding where to go on vacation: Acapulco or Cancún. It turns out that half of your 100 closest friends have been to Acapulco, and the other half to Cancún, but none of them have been to both. You survey these friends to see which place is a better place for a vacation, asking them each to assess how fun their most recent vacation trip to Mexico was, on a scale of one to five.[2]

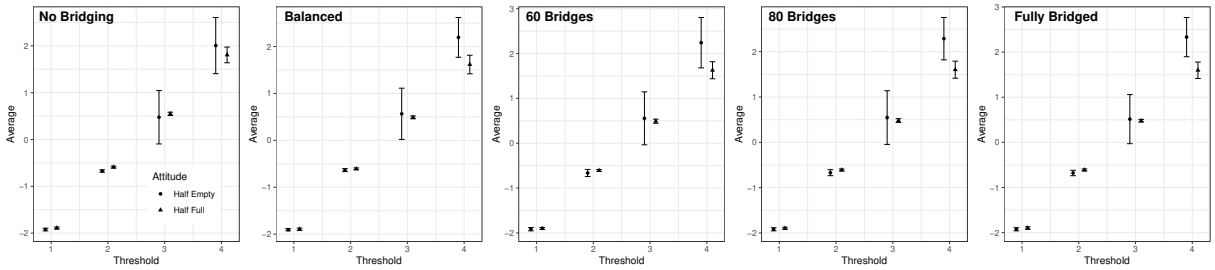
Your friends also fall into two types: glass half-full and glass half-empty. Using the terminology we introduced in the previous chapter, glass half-full people use thresholds that are less strict than glass half-empty people when they translate their perception of fun into survey responses. For this example, we'll assume the glass half-empty people translate their observations down one point on the five-point Likert scale, on average, relative to the half-full folks. We will also assume that the amount of fun someone has on a trip to place i is a function of a "true" latent fun-factor, z_i, and some idiosyncratic error. Further, let's assume that $z_{\text{Acapulco}} = z_{\text{Cancún}} = 1.65$, so that the inherent amount of fun one can have in each place is identical and that trip fun follows a roughly normal distribution across all destinations (so, in this case, Acapulco and Cancún would be around the 95th percentile of fun for all vacation trips).

However, by some quirk of unfortunate fate, your glass half-empty friends all went to Acapulco and the half-full friends went to Cancún. This is a classic bridging problem. If you simply look at the average scores on your survey, you will conclude, incorrectly, that Cancún is going to be more fun than Acapulco, as Cancún's average will be roughly one point higher than Acapulco's. But you will draw the same erroneous conclusion even if you employ the IRT machinery we introduced in the previous chapter. In principle, IRT should be able to adjust for this sort of threshold difference across respondents, but because respondent types sort perfectly across cases, the model cannot learn about these differences. The first panel in Figure 5.1a presents boxplots of the posterior distributions of the latent trait estimates for Acapulco and Cancún obtained by fitting an ordinal IRT model to data simulated from the assumptions we described above. We can use the values described by this boxplot to calculate the posterior probability that Cancún is more fun than Acapulco, and here that value is 1. The reason for this faulty inference, as the first panel in Figure 5.1b shows, is that the model has not learned how the two sets of beachgoers' thresholds differ. Each panel in Figure 5.1b reports averages and 95 percent credible intervals for respondents' estimated ratings thresholds. As we discussed in the previous chapter, these are cutpoints in the latent space that determine where respondents place items on the reported ordinal scale. The first threshold distinguishes between ordinal ratings of 1 and 2, the second between 2 and 3, and so on. Figure 5.1b reports average threshold estimates across respondent type so that we can see whether the model is correctly picking up which friends are half-full and which half-

[2]This might seem crazy, but Dan—Kyle demurs—is pretty sure this is something Brigitte would actually do.



(a) Estimated fun in Acapulco and Cancún



(b) Estimated thresholds for vacationers, with 95 percent credible intervals

Figure 5.1: Latent trait and threshold estimates for the Acapulco/Cancún toy example

empty. For this first model, the threshold distributions are largely centered on the model's priors (given the lack of information necessary to identify these parameters) and overlap across half-full and half-empty respondents.

But now let's assume a slightly different data generating process. Specifically, let's assume that half of the friends of each type went to both places: we have 25 half-full and 25 half-empty respondents for both beach towns. The second panel in Figure 5.1a shows that we now have the information necessary to correctly estimate the relative fun of the two locations: they are statistically indistinguishable. We also are starting to learn which friends are half-full, and which half-empty. If we look at threshold 4 in the second panel of Figure 5.1b, we can see that the average threshold for half-full respondents has dropped, although credible intervals overlap substantially.[3] Of course, you may have noticed that we still have a bridging problem here. The sets of respondents rating each town remains distinct. We get the right answers only because of a fluke: the distribution of types is exactly equal across the two cases. This perfect balance allows us to produce local estimates for both cases that are equal to each other and globally correct. In this case simple averages of respondent scores would also lead us to the correct conclusion. But this is an exception that proves the rule and highlights how crucial bridging can be. It is quite unlikely that one would obtain such perfect balance in most applications, especially in complex real-world situations where *respondents' thresholds reflect multiple underlying factors that are likely to correlate with case.*

[3]Note that, because our friends only assess fun places, we lack information on their other thresholds. We can see that half-full respondents have lower threshold 4s than half-empty ones, but we still assume that their remaining thresholds are similar to each other, as the priors provide the only information relevant to their estimation.

We will wrap up this toy example by adding some bridging. Specifically we will go back to the original data-generating process—perfect separation by type—but then sequentially add sixty, eighty, and one hundred bridges. These scenarios simulate situations where some portion of our friends visited both resort towns, and we assume an equal number of bridges from the half-full and half-empty groups in each case.[4] The remaining panels in Figure 5.1 show us what happens as we add bridges. With sixty bridges, the estimated distributions for the latent trait are much closer together—Cancún no longer seems dramatically more fun than Acapulco—but there remains a 97 percent probability that Cancún is *more* fun. With eighty bridges, that probability drops to 70 percent, a statistically insignificant difference if one applies frequentist norms to our Bayesian probability estimates. With full bridging, we correctly estimate that the trips are a toss-up for fun and the probability that Cancún is more fun is now 47 percent. The final panel in Figure 5.1b also shows that the model is now quite sure that the half-empty respondents have a higher top threshold than the half-full ones, accurately reflecting their increased standards. Results slowly converge toward reality as we add bridges.

This example turns out to be a rather difficult case. We have very limited information about respondent thresholds, even within subgroups, because we receive only one or two—once we add bridges—ratings from each friend. As a result, we need substantial bridging to ascertain that Acapulco and Cancún are equally fun. In other contexts, as we shall see, it is possible to obtain accurate estimates with far fewer bridges. Indeed, this ability to work with sparse bridging is one of the major advantages of the latent variable modeling framework that we present throughout this book. By using the tools of Chapter 4 to make respondents comparable, we can derive accurate measurements across cases, even when the majority of the respondent pool exhibits crucial—and case-correlated—differences.

We can also use our animal-beatability survey to illustrate the bridging problem more concretely with real data. In addition to asking respondents which animals they thought they could beat in a fight, we asked respondents to evaluate how "typical" people would do in such contests. We recruited approximately 90 respondents each from Australia, the United Kingdom, and the United States and asked each one to assess fifteen animals on a single binary question: could a given target person beat each animal in an unarmed fight? Crucially, each respondent answered this question for three distinct target *cases*: the average American (Q16), the average Briton (Q17), and the average Australian (Q18). The structure maps directly onto a cross-case expert survey: the fifteen animals are the items, the three national averages are the cases, and the respondents are the raters. A home rater is an American rating the average American, a Briton rating the average Briton, and so on.

Now suppose we proceeded as most cross-case surveys do and used only home raters—Americans to characterize the average American, Britons the average Briton, and Australians the average Australian. We established in Chapter 4 that these respondents bring systematically different response thresholds in their evaluations of their own ability to defeat different animals. Australians, who may encounter more dangerous wildlife on a regular basis, tend to set higher bars for what counts as a beatable animal. Britons occupy an intermediate position. Great Britain's megafauna are largely gone, so British respondents have extensive experience with the same common animals that everyone has seen (rats, cats, geese, and

[4]So the sixty bridges are provided by thirty half-full and thirty half-empty respondents, and so on.

dogs) but little direct exposure to the distinctive megafauna of either Australia or North America. Americans, drawing on experiences with megafauna that include large predators but few of the hazards routine to Australian life, tend to be more generous raters. Those threshold differences do not vanish when raters shift from assessing themselves to assessing their national average. An Australian who self-rates conservatively is likely to rate the average Australian conservatively. Without bridge respondents—people who rate more than one national average—the three scales float freely relative to each other. The apparent gaps among national averages could reflect genuine differences in fighting ability or simply differences in national rating tendencies, and we have no way to tell them apart.

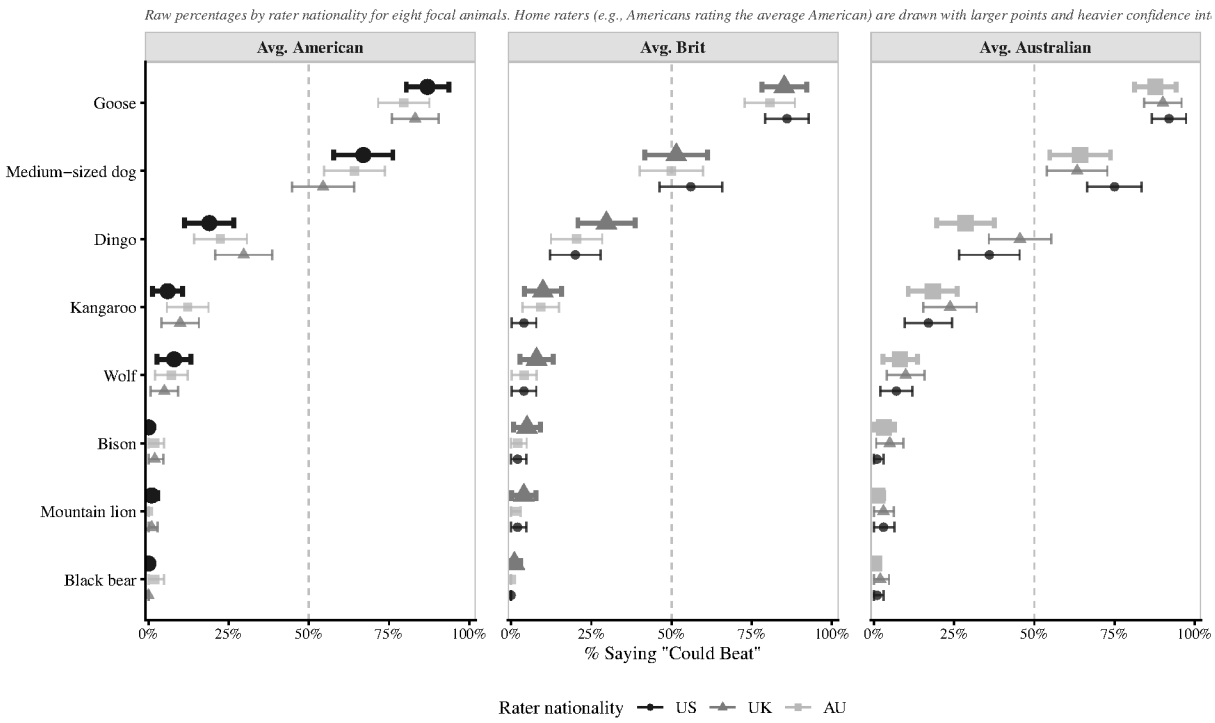
Figure 5.2 illustrates this dilemma. Each panel shows, for one of the three national-average cases, the raw percentage of each rater group—American, British, and Australian—who believe that national average could defeat a focal animal; home raters are distinguished from out-of-case raters. There are clear cross-country differences in evaluations which highlight two potential underlying causes of cross-case DIF.

The first is direct experience. Consider the dingo, which is native to Australia. Australian home raters are relatively pessimistic about the average Australian's dingo odds (29 percent), whereas British and American raters are considerably more sanguine (46 percent and 36 percent, respectively). Australians know what dingoes are actually like—fast, cooperative hunters built for coursing prey—and they adjust accordingly. Britons and Americans, who have never encountered a dingo outside of a nature documentary, tend to treat it as roughly equivalent to a mid-sized wild dog. A similar dynamic appears for bison. American home raters give the average American essentially no chance against a bison (0 percent), as they are uniquely aware that it is responsible for more visitor injuries in US national parks than bears, mountain lions, and wolves combined. British and Australian raters, for whom the bison is a large but seemingly docile bovine of distant plains, assign a small but real probability of a human winning the fight (2 percent). In both cases, the home raters' direct exposure produces a conservatism that out-of-case raters, reasoning from general principles, cannot replicate.

The second factor runs in the opposite direction: stereotype. When Britons and Americans assess the average Australian across the full range of animals in the survey, they are systematically more generous than Australians are in their self-assessments. The gap is largest for the dingo, where British raters put the average Australian's odds at 46 percent—17 percentage points above what Australian home raters report. But this difference spans the profile: across all focal animals, UK and US raters are 3 to 4 percentage points more optimistic about the average Australian than Australians are about themselves. The cultural archetype of the tough, wildlife-hardened Australian—think Steve Irwin or Crocodile Dundee—might explain this regularity, inflating out-of-case assessments beyond what home raters recognize as warranted. And stereotypes can cut in more than one direction. Because Australians know how dangerous dingoes actually are, they rate the average Briton's odds against a dingo at only 20 percent, while British home raters put themselves at 30 percent.

This is, in miniature, precisely the bridging problem. The average Australian looks one way to Australians and another to everyone else; the average Brit looks one way to Britons and another to Australians who know what they do not know. Some of the divergence reflects genuine informational advantages; some reflects cultural lenses that have nothing to do with the underlying reality. Without a measurement framework that can separate the

Figure 5.2: Raw percentage responding "Could Beat" for eight focal animals, by rater nationality and case. Each panel corresponds to one national-average case; each point is the mean percentage of a rater group who believe that case could defeat the focal animal (with 95 percent binomial confidence intervals). Home raters (e.g., Americans rating the average American) are drawn with larger points and heavier confidence intervals, and out-of-case raters with smaller, lighter markers. Animals are ordered by overall mean across all raters and cases.



two, and without bridge respondents who rate more than one case and thereby connect the scales, there is no way to know which assessments to trust.

5.2 Bridging Different Sorts of Surveys

5.2.1 Expert surveys

While our simulated Acapulco and Cancún trip ratings from earlier in the chapter provide a simple window into the bridging problem and illustrate its key features, expert surveys represent a real-world example of how social scientists have long ignored bridging issues in survey research. Social scientists typically employ expert surveys when they lack easily observable proxy variables for critical concepts, but when case experts have the information necessary to provide subjective assessments of how cases relate to concepts of interest. Such surveys would traditionally aggregate information from a set of experts on each relevant

case—in our field of comparative politics, typically a country—and ask them to answer survey questions focused solely on the case for which they were recruited. For example, comparative scholars of political parties have long used expert surveys to estimate the ideological positions of parties around the world (Castles & Mair 1984, Huber & Inglehart 1995, Laver & Hunt 1992, Budge 2000), but only in recent years have such surveys used bridging techniques, such as anchoring vignettes, to help ensure that experts understand and apply concepts similarly across countries (Bakker, Jolly, Polk & Poole 2014).

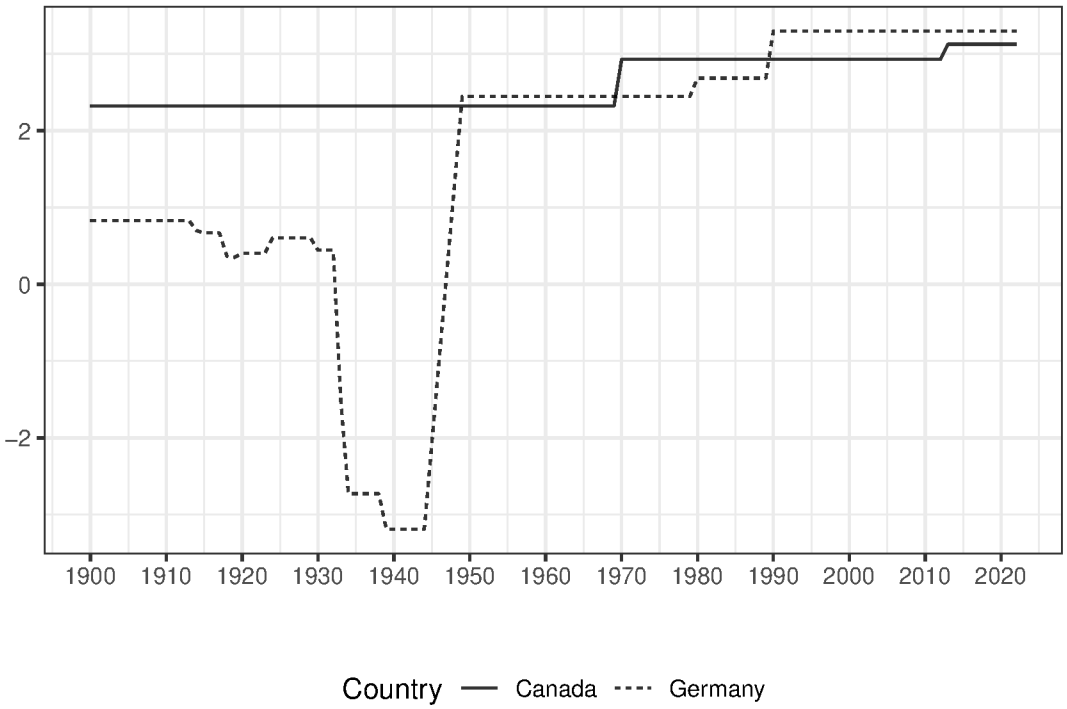
Similarly, expert surveys of social institutions have proliferated in recent years (Cooley & Snyder 2015). These projects measure everything from corruption (Transparency International 2023) to good governance (Anheier, Lang & Knudsen 2022), electoral integrity (Garnett, James & MacGregor 2022), state capacity (Hanson & Sigman 2021), and quality of democracy (Coppedge, Gerring, Knutsen, Lindberg, Teorell, Altman, Bernhard, Cornell, Fish, Gastaldi, Gjerlow, Glynn, Grahn, Hicken, Kinzelbach, Marquardt, McMann, Mechkova, Neundorf, Paxton, Pemstein, Rydén, von Römer, Seim, Sigman, Skaaning, Staton, Sundström, Tzelgov, Uberti, Wang, Wig & Ziblatt 2023). And party ideology measurement projects continue to emerge (Lindberg, Düpont, Higashijima, Kavasoglu, Marquardt, Bernhard, Holger, Hicken, Laebens, Medzihorsky, Neundorf, Reuter, Ruth-Lovell, Weghorst, Wiesehomeier, Wright, Alizada, Bederke, Gastaldi, Grahn, Hindle, Ilchenko, von Römer, Wilson, Pemstein & Seim 2022, Norris 2019). Yet, few of even these more recent surveys attack the bridging problem head-on. Rather, most ask respondents to assess only one country and report means and standard deviations across respondents, at the case-question level.

In principle, asking respondents to assess only the cases for which they are most expert makes sense. One of the primary reasons that one recruits experts for such surveys is that they know things about their cases that other people do not. They often may lack such information for any case beyond their specialized case. Nonetheless, the one-case-per-expert approach essentially assumes that expert thresholds are identical, regardless of which case they evaluate, and that all experts perceive the latent concept identically, or at worst with random error. In some contexts, these assumptions may be justified. For instance, Bakker et al. (2014) use anchoring vignettes and Aldrich-McKelvey scaling[5] to bridge expert-based party ideology scores from the Chapel Hill Expert Survey (CHES) and show that simple respondent means correlate highly with bridged estimates, generally validating the assumption in the context of estimating European political parties' ideological positions. In other words, in this specific context, the researchers have demonstrated that there is no bridging problem. Nonetheless, it was by explicitly acquiring bridges, and assessing the extent of the potential problem, that these authors were able to validate a widely used assumption—but that assumption could have proven wrong.

Assuming such equivalence across experts and cases is a leap of faith, and one that can lead to puzzling results. For instance, the Electoral Integrity Project famously rated the quality of democracy in North Carolina as equivalent to that in Cuba, setting off a flurry of skeptical news reports and rebuttals (Gelman 2017). This puzzling equivalence could reflect a variety of sources of bias. The most likely explanation, however, is that the Electoral

[5]Aldrich-McKelvey scaling is a technique for placing survey respondents' ratings of political stimuli on a common scale by estimating, and adjusting for, each respondent's individual distortion—shift and stretch—of the underlying dimension.

Figure 5.3: V-Dem estimates of "Freedom from political killings"



Integrity Project expert who was rating the quality of elections in the established democracy of North Carolina was operating from a different reference point than the expert who was rating the autocracy of Cuba, where elections play a qualitatively different role. In other words, these respondents applied different standards for excellence in election integrity to two very different cases—their thresholds were fundamentally different.

Here we will use the V-Dem dataset to illustrate some real-world bridging problems, and to highlight how bridging cross-case respondents can be essential to effective measurement in cross-national expert surveys. The V-Dem data provide a fertile resource for evaluating the extent to which bridging problems afflict the cross-national rating process because of their breadth and depth—they assess many cases across many variables—but also because they include multiple bridging observations. V-Dem employs experts who respond to multiple cases and uses anchoring vignettes to see how their respondents assess standardized hypothetical situations. We can see what V-Dem estimates would look like without these bridging observations by removing them from the dataset before estimating country scores. As we shall see, while country time-series remain internally consistent after such an adjustment, the relative ratings of different countries can fall substantially out of whack.

Figure 5.3 shows time series of V-Dem's "Freedom from political killings" variable for Canada and Germany, from 1900 through 2022. The figure, which one can generate using V-Dem's online graphing tools,[6] plots latent trait estimates over time. The chart helpfully

[6]See `https://V-Dem.net/graphing/graphing-tools/`.

includes estimates of the breakpoints for scores on the ordinal scale that experts used (marking them with gray lines) and labels codebook categories on the y-axis.[7] Canada exhibits freedom from political killings near the top of the scale across the entire period, climbing a touch over time. Germany, on the other hand, shows a somewhat lower level of freedom from killings in the lead-up to the Second World War, a very large drop while the Nazis were in control, and then a quick rise to a level similar to Canada's in the post-war period (they are statistically difficult to distinguish after 1948). Two pools of experts generated the data that informed these estimates. Seventeen experts provided responses for Canada, while twenty-six assessed Germany. However, these experts assessed different spans of time. Four Canadian experts assessed the full time-series, while two others rated most of it. One respondent assessed only the period from 1900 to 1920,[8] and ten respondents assessed only portions of the twenty-first century. Eleven German experts assessed the entire time-span, ten the period from 1900 to 1920,[9] and five focused on the twenty-first century.

One can bridge expert surveys using multiple methods and V-Dem has obtained bridging information using three different approaches (Coppedge, Gerring, Knutsen, Lindberg, Teorell, Marquardt, Medzihorsky, Pemstein, Pernes, von Römer, Stepanova, Tzelgov, Wang & Wilson 2019). First, V-Dem asked individual experts to do what is known within the project as *bridging*. Bridging requires an expert for one country to complete the survey for another country. When possible, such bridge respondents complete full time-series for multiple countries, but partial bridging—where a respondent completes part of the time series for more than one country—is also useful for anchoring scores across cases. This approach may be the ideal way to achieve bridging, as the resulting data provides substantial information about both cases and respondents. It also has the benefit of asking experts to evaluate real cases. And it can, in principle, provide reasonable coverage. While there are typically few experts who can assess many cases, if many respondents can assess a few cases each, then a survey researcher can obtain a dense overall bridging network. But this way of obtaining bridges can be costly and time-consuming. Respondents with limited time may be unlikely to agree to assess a substantial number of bridges for longer surveys, even when they have the capability.

The second approach involves certain respondents that have wide-ranging expertise. For example, in multi-country studies, regional experts might be able to assess cases across a reasonably large set of countries, such as every country on a given continent. When surveys are short, it might make sense for such respondents to bridge multiple cases fully. But when surveys are long—either because there are many questions, or because the respondent assesses each case repeatedly over time, or both—it may make more sense to ask such wide-ranging experts to respond to the full survey for one case and parts of the survey for other cases. In the V-Dem survey, for example, multiple regional experts engaged in what the project calls "lateral coding"—rating up to fourteen countries beyond their primary case for a single year. This sort of bridging helps to reduce the number of "islands" and "continents"

[7] V-Dem uses a hierarchical IRT framework that allows one to estimate "global thresholds" across respondents. These equate roughly to the average respondent thresholds.

[8] This was an Historical V-Dem respondent who assessed a twenty-year overlap with the contemporary period.

[9] There were a lot of German principalities in the historical period, so many "historical experts" provided contemporary bridges.

in the bridging network (see section 5.4.1), but produces less detailed information about latent traits, and respondent threshold parameters, than full-scale case bridging.

Finally, the third approach involves using anchoring vignettes in expert surveys (King & Wand 2007) to provide bridging when expertise is truly case-specific, so that respondents effectively cannot provide real bridges. Anchoring vignettes are also a nice option when respondents can technically assess cases outside their core expertise, but such work is too time-consuming or costly. They can also be deployed—as V-Dem does—to supplement existing bridges. Anchoring vignettes are simply short descriptions of hypothetical cases which respondents assess relative to the same survey questions that they answer about real cases.[10] In principle, such vignettes provide the perfect tool for bridging. Any expert can assess them, they are reasonably cheap to produce, and a researcher can develop hypothetical cases that span the latent space, allowing one to learn about respondent's thresholds efficiently. Nonetheless, these anchoring vignettes have a number of drawbacks. First, it can be difficult to write vignettes well. In principle, respondents should rank vignettes consistently, even if they give them somewhat different ratings. But writing such consistently behaved vignettes can be quite difficult in practice and one generally needs to test many potential vignettes before settling on a subset to deploy. Limited space on a survey makes it difficult to provide rich descriptions that approximate real cases, and respondents may evaluate vignettes differently—essentially use a different latent scale—than they do real cases.

Figure 5.3 displays results of a measurement model that uses information drawn from all three bridging strategies. First, multiple respondents assessed countries outside of Canada and Germany. The estimates in Figure 5.3 are, therefore, influenced by a complex network of "natural" bridges that allow the model to learn where Canadian and German experts' thresholds sit relative to others. Many of these experts also responded to anchoring vignettes for this question, which provide dense "synthetic" bridges that provide substantial information about respondent thresholds.

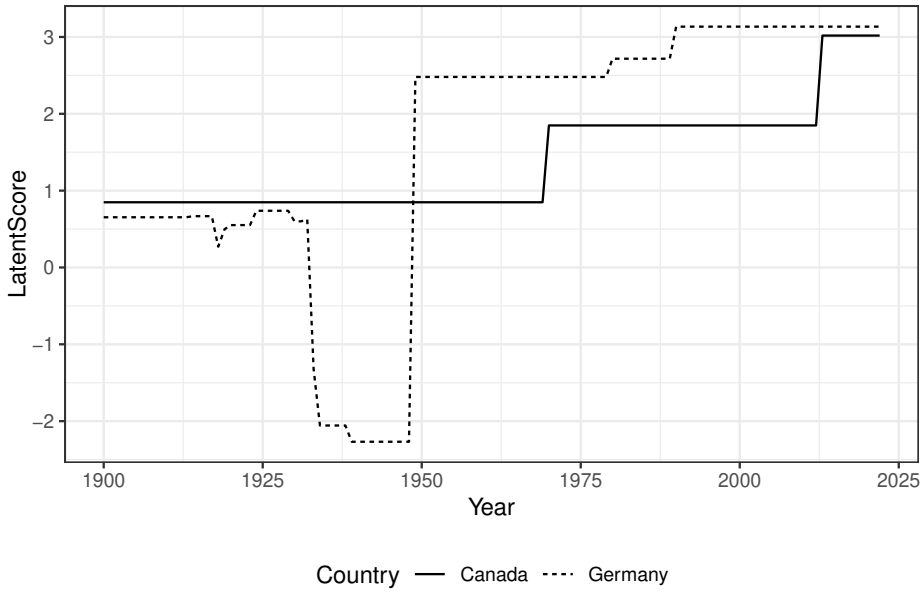
Figure 5.4 shows what happens when we remove these bridges. We estimated a simple ordinal IRT with vague priors for threshold parameters,[11] using only the data for Canada and Germany, eliminating all bridging. From a within-country perspective, the time series look almost identical to the official V-Dem measures.[12] But the cross-national comparison has changed. While the estimates for Canada are generally above or equal to Germany's according to V-Dem's official—bridged—estimates, our unbridged model has pushed Canada's estimates down until the very end of the period under study. The reason for this is simple. The model lacks the information necessary to place thresholds for Canadian and German experts on the same scale, and must rely instead on priors to set those thresholds. This issue is exacerbated by the fact that most Canadian respondents provide routinely high values and therefore provide the model with virtually no information about their lower thresholds. Lacking information about Canadian thresholds, and with the prior for the latent variable itself centered on zero, the model will tend to push relatively constant time-series toward

[10] We discuss anchoring vignettes in more detail in the next section, as well as in Chapter 6.

[11] We adopt standard $\mathcal{N}(0,1)$ priors on the latent traits, $\mathcal{N}(1,1)$ priors—truncated below at 0—for the discrimination parameters, and the four threshold parameters are drawn from the prior $\mathcal{N}(\{-1.5,-0.5,0.5,1.5\},2)$, subject to an ordering constraint.

[12] Remaining differences likely stem from subtle differences in how we collapse yearly data into ratings regimes (Pemstein et al. 2023). See chapter 7 for a discussion of how, and why, V-Dem collapses yearly data.

Figure 5.4: Unbridged estimates of "Freedom from political killings"



the center of the latent scale. We can only extract sensible cross-national estimates by introducing bridges and helping the model to learn the relative placement of individual experts' thresholds.

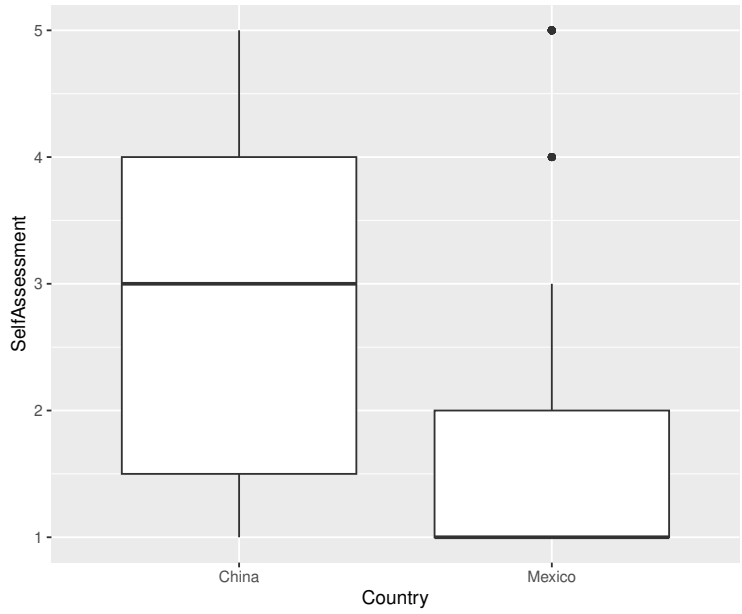
5.2.2 Layperson surveys

While expert surveys ask respondents to assess things that are external to them but which they know a lot about, layperson surveys mostly focus on eliciting information about respondents' personal beliefs or experiences. This difference in focus will tend to reduce the bridging options for layperson surveys relative to expert surveys, because lay respondents will rarely have the capacity to assess multiple cases. By definition, respondents who provide information about their personal opinions, feelings, or circumstances simply cannot assess any case but their own.

As our discussion at the outset of this chapter highlighted, layperson surveys can also be subject to substantial bridging issues. As with expert surveys, cross-country research presents an obvious potential for problems with scale equivalence across cases.[13] For instance, King et al. (2004) examine a bridging problem with a World Health Organization (WHO)

[13]Scale equivalence can also be an issue within cases, of course. For example, even a simple presidential approval survey in one country at one point in time can produce misleading inferences if respondents' scale thresholds correlate with their assessments. For instance, say the sorts of respondents who are unhappy with the president systematically use higher response categories to report their feelings. This example stretches the bridge problems concept, but it is worth noting that bridging approaches that we discuss here, and the toolkit developed in Chapter 4, can be useful even when it is difficult to assign respondents to discrete cases or groups, ex ante.

Figure 5.5: Distribution of self-assessed political efficacy by Chinese and Mexican respondents



survey that asked people in China and Mexico questions about political efficacy, among other things. Specifically, they asked respondents:

> How much say do you have in getting the government to address issues that interest you? (1) No say, (2) Little say, (3) Some say, (4) A lot of say, (5) Unlimited say.

King et al. (2004) note a conundrum with this question. Chinese respondents—who live in an autocracy—report systematically higher political efficacy than Mexican respondents—who live in a democracy. But a strong case can be made that Mexican respondents have objectively higher political efficacy, as they have substantially more say in governing through voting and political freedoms than Chinese citizens have. This unexpected result is consistent with cross-case differences in how respondents understand, and respond to, the WHO's survey question. Figure 5.5 displays a box-plot of the self-assessments provided by Chinese and Mexican respondents, illustrating the bridging problem faced here.

Luckily, King et al. (2004) added *bridging observations* to address this problem. In this context, one cannot obtain explicit cross-case bridges. That is, it is impossible for a Mexican respondent to assess the level of political efficacy that she would have if she lived in China, and vice versa. But such respondents can answer questions about how they would assess the political efficacy of hypothetical people, each imagined as living in the respondent's own country, simulating the process of explicit bridging. Specifically, they asked both the Chinese and Mexican respondents to assess the political efficacy of hypothetical people, including these examples:

Case	Response 1	Response 2	...	Response $N-1$	Response N
China	NA	NA	...	5	4
Mexico	1	1	...	NA	NA
Alison	5	1	...	5	5
Jane	5	5	...	3	3
Moses	2	5	...	5	4

Table 5.1: Organizing the WHO data for IRT analysis.

Alison lacks clean drinking water. She and her neighbors are supporting an opposition candidate in the forthcoming elections that has promised to address the issue. It appears that so many people in her area feel the same way that the opposition candidate will defeat the incumbent representative.

Jane lacks clean drinking water because the government is pursuing an industrial development plan. In the campaign for an upcoming election, an opposition party has promised to address the issue, but she feels it would be futile to vote for the opposition since the government is certain to win.

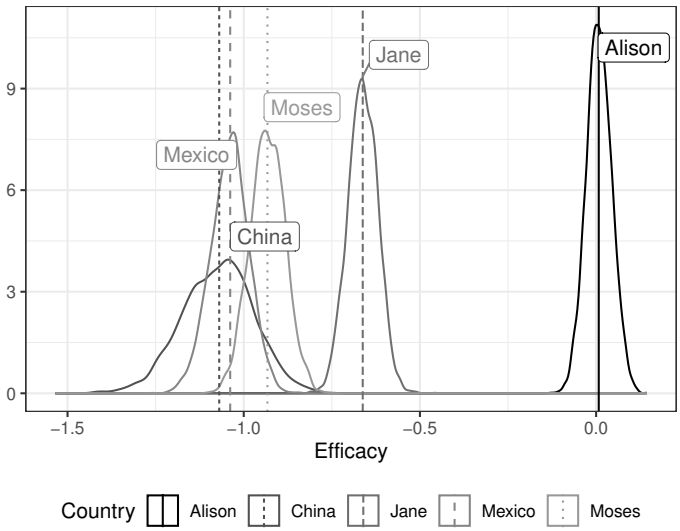
Moses lacks clean drinking water. He would like to change this, but he can't vote and feels that no one in the government cares about this issue. So he suffers in silence, hoping something will be done in the future.

Using these sorts of *anchoring vignettes* as bridges is the topic of Chapter 6, but we preview their use here to illustrate how our IRT framework can be applied to address cross-case surveys of regular people when anchoring vignettes are included in the survey, and to highlight the fact that such vignettes are a form of *bridging observation*. First, we organize the data,[14] as Table 5.1 shows, with cases on the rows and respondents on the columns. While every respondent provides a rating for the hypothetical Alison, Jane, and Moses, we have missing data for the Mexico and China cases, because each respondent only provides their self-assessment with respect to their own country.

Figure 5.6 shows the results of fitting an ordinal IRT to these data—including not just the self-assessment, but also each respondent's evaluation of the three hypothetical people's political efficacy as observations. The figure provides density plots of the estimated posterior distributions of the latent traits for each case, both the real cases of China and Mexico, and the hypothetical bridging cases of Alison, Jane, and Moses. We can see that after incorporating the bridging information available in the anchoring vignettes into our model, we estimate the political efficacy of the typical Chinese respondent to be slightly less than that of the typical Mexican respondent, although the posterior distributions overlap substantially. We can also see that the respondents assess the anchoring vignettes as we would expect: Alison has substantial, Jane has middling, and Moses has low political efficacy. We might

[14]The data that we analyze in this section come from the supplementary materials for Imai & Webb Williams (2022), exercise 3.10.2, and can be found at `https://github.com/kosukeimai/qss/blob/master/MEASUREMENT/vignettes.csv`.

Figure 5.6: Ordinal IRT estimates of the posterior distributions of latent efficacy scores for China, Mexico, and Vignettes



Note: Vertical lines are median latent efficacy estimates.

still question whether Mexicans truly have essentially the same level of political efficacy as Chinese respondents, as the results imply. But given the bridging information, this similarity in self-assessment stems from the fact that both sets of respondents see their own political efficacy, on average, as similar to—or even lower than—Moses's. We have used our bridges to place each case on the same scale.

The model's estimated gamma parameters also illustrate that the Mexican respondents have higher standards than the Chinese ones. For Mexico, the average estimated threshold scores[15] on the latent scale are {-1.6, -0.4, 0.6, 1.7}. For China, the corresponding average thresholds are {-2.1, -1.0, 0.1, 1.3}. Thus, for the typical Mexican respondent, the difference between "No Say" and "Little Say" in government falls at -1.6 on the latent scale, while for the typical Chinese respondent, it is far lower at -2.1. This pattern is consistent across the latent scale, as Chinese respondents have lower *thresholds* than Mexican respondents. Or, as King & Wand (2007), who also re-analyze these data, conclude "[T]he Chinese report higher *levels* of say in government because they have lower *standards* for what counts as satisfying the level described by any given response category."

Anchoring vignettes are generally the go-to tool for bridging layperson surveys, because they work in situations where a researcher is interested in measuring individuals' assessments of their personal situations, feelings, or beliefs. A respondent need not know anything about how anyone else feels, or what they have experienced, to evaluate an anchoring vignette. As long as respondents can assess hypothetical states of affairs, mapping theoretical experiences that differ from their own onto the survey question's response scale, then vignettes can serve as bridges. Of course, as we noted in the previous section, and as Chapter 6 further

[15]Averaged across both respondents and posterior draws.

explicates, effectively deploying vignettes requires great care, and their value rests on crucial assumptions about consistency between real and hypothetical assessments.

5.2.3 Ideological common spaces

One context where bridging observations play a key role is in the estimation of political ideology. A number of scholars have attempted to estimate ideological positions of political actors (e.g., how liberal or conservative they are) in latent space by creating "survey items" using voting records or other expressed preferences about policy options.[16]

Current work largely builds on Clinton, Jackman & Rivers's (2004) insight that the spatial voting model with quadratic utility functions—central to a plethora of game theoretical works on voting behavior—is mathematically equivalent to a two-parameter IRT model when one adds random errors. These spatial models conceptualize policy outcomes and actors' preferences over those outcomes as points in some K-dimensional latent space.[17] In other words, the models assume that we can represent votes in terms of two points in this latent space. Specifically, voters must choose between the "Yea" position $\boldsymbol{\zeta_j}$ and the "Nay" outcome $\boldsymbol{\psi_j}$ on each of $j \in 1 \dots m$ votes. Each voter has a preferred policy, or *ideal point*, $\mathbf{x_i}$.[18] Voting models assume that voters' utilities for a given policy outcome are functions of the distance between the policy and their ideal points, and the standard model assumes the loss function is quadratic, largely because it makes the math easy. To turn the game theoretic setup into a statistical model of voting behavior, we simply add error terms and assume that legislator $i \in 1 \dots n$ with ideal point $\mathbf{x_i}$ derives expected utility

$$U_i(\boldsymbol{\zeta_j}) = -||\mathbf{x_i} - \boldsymbol{\zeta_j}||^2 + \eta_{ij} \tag{5.1}$$

from voting yea on vote j, and

$$U_i(\boldsymbol{\psi_j}) = -||\mathbf{x_i} - \boldsymbol{\psi_j}||^2 + \nu_{ij} \tag{5.2}$$

from voting to reject. Utility-maximizing voters vote yes ($y_{ij} = 1$) when $U_i(\boldsymbol{\zeta_j}) > U_i(\boldsymbol{\psi_j})$, and no ($y_{ij} = 0$) otherwise.

If we assume that η_{ij} and ν_{ij} are independent with respect to both voters and votes, and normally and jointly distributed with mean $E(\eta_{ij}) = E(\nu_{ij})$ and variance $\text{var}(\eta_{ij} - \nu_{ij}) = \sigma_j^2$, then $P(y_{ij} = 1) = P(U_i(\boldsymbol{\zeta_j}) > U_i(\boldsymbol{\psi_j}))$. We can rewrite this equation as

$$P(y_{ij} = 1) = \Phi\left(\boldsymbol{\beta_j}(\mathbf{x_i} - \boldsymbol{\kappa_i})\right), \tag{5.3}$$

where $\boldsymbol{\kappa_j} = \frac{\boldsymbol{\psi_j} + \boldsymbol{\zeta_j}}{2}$ is the cutpoint dividing voters who support measure j from those who do not and $\boldsymbol{\beta_j} = \frac{2(\boldsymbol{\zeta_j} - \boldsymbol{\psi_j})}{\sigma_j}$ describes the extent to which vote j discriminates between voters' baseline policy preferences.

[16] See Poole (2005) for an overview or approaches to using measurement models to estimate ideology in legislatures. Martin & Quinn (2002) apply similar techniques to Supreme Court decisions, while Voeten (2000) uses votes in the United Nations General Assembly to generate an ideological map of states' positions on international relations.

[17] Typically, K is one or two, since humans have difficulty visualizing higher-dimensional spaces. It also turns out that such low-dimensional models predict the lion's share of variance in actual voting behavior, across most domains.

[18] Note that these parameters are K-vectors, representing points in K-space.

You may recognize this equation as a version of the ordinal IRT model we have relied upon throughout this book, although here there is only one threshold (the cutpoint between yes- and no-voters) to estimate. The discrimination parameter, $\boldsymbol{\beta_j}$ is, as in our workhorse model, a measure of how informative a given vote is about voters' ideal points. The only major difference between this model and our workhorse model is that our focus has been on one-dimensional latent traits, while this model allows for any number of dimensions in the latent policy space.[19] Also note that while the model is typically expressed in terms of "votes," one can apply it whenever respondents choose between two outcomes. A researcher can apply it to either-or survey questions about policy preferences or to any binary choice of interest.

Much of the scholarship that uses this model focuses on a singular institution, such as a legislature or a court, and uses the patterns of votes in that body to draw inferences about where individual legislators or justices sit in ideological space. Bridging becomes an issue when a researcher wants to fit the model to a given set of people (Congress, the Supreme Court, or a random sample of voters) across multiple time periods, or to compare ideology across different types of respondents. For instance, the composition of legislatures varies across terms, and certain legislators (e.g., one who retires in term $t-1$ and another who starts their career in the legislature in term t) will never vote on the same piece of legislation. But since incumbents tend to be re-elected, there are normally enough legislators to bridge those parliamentarians who never vote on the same set of bills.

Things become much trickier when one wants to produce comparable estimates of ideology across different institutions, or across elites and members of the voting public. For instance, Bailey (2007) attempts to generate comparable policy preference estimates across US members of Congress, presidents, and Supreme Court justices. This setup presents essentially the same problem that V-Dem faces when attempting to estimate freedom from political killings based on the responses of Canadian and German experts, or when the WHO tries to generate comparable estimates of Chinese and Mexican respondents' feelings of political efficacy. Bailey (2007) draws on a variety of sources to create bridges between institutions. First, he uses presidents' statements on Supreme Court cases, coding them as pro- or anti-decision, to create a set of "synthetic votes" by presidents on those cases. Because these are somewhat rare, he augments presidents' statements with "votes" drawn from solicitors general's amicus briefs. Bailey assumes that each solicitor general, appointed by the president, represents the administration and, in turn, the president's own point of view. Presidents' statements on Senate and House votes serve as similar bridges between the president and Congress. He also builds bridges from the House and Senate to the Supreme Court by coding statements on Supreme Court decisions in the *Congressional Record*, Congressional representatives' amicus briefs, and sponsorship of—and roll call votes on—legislation specifically taking a position on Supreme Court cases. Finally, he augments the Supreme Court's voting record by looking for phrases like "wrongly decided" in Court opinions and codes votes in line with those opinions as votes for or against prior cases to help bridge decisions across courts.

While, at first glance, common-space voting models are quite distinct from the more clear-cut surveys we have discussed so far, they provide a useful example of how bridging

[19] IRT models generalize to multiple dimension by treating latent trait parameters as vectors instead of scalars. These sorts of multi-dimensional IRT models are beyond the scope of this book.

problems can plague measurement in social research. They also show how, if one is willing to make reasonably strong assumptions, one can find bridges in circumstances where both true bridging and hypothetical assessments are not possible. In the example drawn from Bailey's research on the three branches of the US federal system, assuming that presidential statements provide the same sort of information as official legislative votes facilitates the creation of a bridge between disparate survey information. Regardless of the modality of the "survey" in question, making clear assumptions about these sorts of quasi-bridges may often be more desirable than just assuming the bridging problem away.

5.2.4 Ratings sites

As we stressed in the introduction, bridging problems are a huge concern for the ratings that consumers give to goods and services on ratings sites. Recall the example of Yelp ratings for Chinese restaurants in Fargo, North Dakota and Champaign, Illinois, from the Introduction. Both towns have 4.5-star restaurants—Mandarin Express in Fargo and Golden Harbor in Champaign—but only someone who has been to both of these towns, and had a hankering for Chinese food while in both places, is likely to understand how drastically out-of-sync these two ratings are. Both towns are in largely rural areas, far from the large immigrant communities in urban areas that normally produce top-notch Chinese fare, but Champaign has lucked out while Fargo, alas, has not. This one example illustrates a general problem. Ratings-site product evaluations almost never take DIF, or bridging problems, into account, and we often need to use context cues to adjust such evaluations for ourselves. In other words, as most of us realize, any five-star review is inherently context-specific. If we take these reviews at face value, we assume that the people providing the stars have identical thresholds (no DIF) and that the distribution of rater thresholds is the same across places (good bridging). Most travelers intuitively understand that they need to guestimate local standards and situate those standards in alignment with their own in order to effectively use such ratings to pick a restaurant.

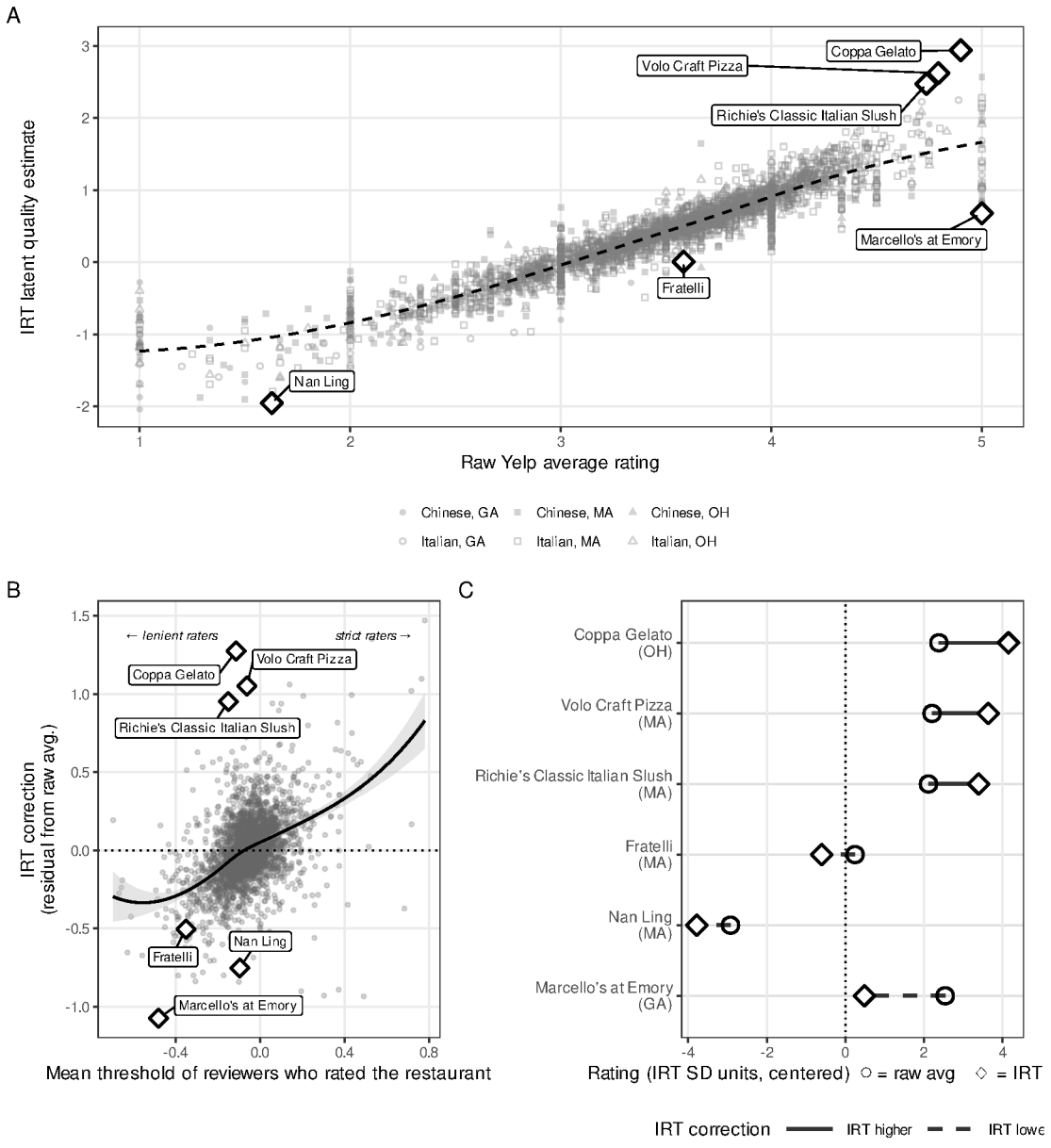
Fortunately, we can use the tools that we develop in this book to overcome many of the ratings inconsistency issues that one faces when consulting reviews from ratings sites. Here, we use Yelp's open dataset[20] to illustrate how to apply our IRT framework to ratings-site reviews. We decided to focus on Chinese and Italian restaurants in Georgia, Massachusetts, and Ohio, as the Yelp open dataset provides substantial coverage in these three markets. We focus on two specific cuisines, rather than a random subset of restaurants, because concentrating on common cuisine categories produces denser cross-state bridging: a traveler is more likely to have reviewed, say, Chinese food in more than one state. To speed up the estimation process, and to ensure a sufficiently dense ratings matrix, we dropped reviews by users who evaluated fewer than five of the restaurants. The resulting data represent 13,615 respondents reviewing 3,093 different restaurants for a total of 138,101 reviews.

We fit an ordinal IRT model to the data, using vague priors.[21] Panel A in Figure 5.7 plots average ratings against point estimates of latent restaurant quality estimated by the model. While estimated quality tracks raw averages overall, many scores are shifted substantially

[20] You can get these data at `https://business.yelp.com/data/resources/open-dataset/`.

[21] We used standard normal priors for the latent traits, $\mathcal{N}(1,1)$ priors for the discrimination parameters, and $\mathcal{N}(\mathbf{x},1)$ priors for each of the threshold parameters, where $\mathbf{x} = \{-1.5, -0.5, 0.5, 1.5\}$.

Figure 5.7: Three views of ordinal IRT estimates of restaurant quality for Chinese and Italian restaurants in Georgia, Massachusetts, and Ohio. *Panel A*: IRT latent quality estimates versus raw Yelp average ratings. *Panel B*: IRT correction plotted against reviewers' leniency, defined as average estimated threshold parameters. *Panel C*: Before-and-after comparison for the six restaurants with the largest corrections. The same six restaurants are highlighted with open diamonds in Panels A and B.



away from the 45-degree line. This implies that raw averages, which make no use of bridging information, are telling a different story than our approach, which takes advantage of the bridging information that we have. For instance, shifts are especially prominent at integer values along the x-axis. These are often cases with only one rating. One big issue with these sorts of cases is that it is difficult to know when an assessment produced by only a few respondents can be trusted. But our framework takes advantage of bridging to place such reviews in context.

Consider Marcello's at Emory, an Atlanta-area Italian restaurant for which the open dataset contains exactly one (five-star) review. There are more reviews that appear online, but they are not included in the open dataset. The IRT model assigns this restaurant the lowest estimated quality—0.7, with a one standard deviation credible interval ranging from -0.2 to 1.5—of any five-star restaurant in the dataset. This is a typical score for restaurants with about 3.8 stars, although the credible interval covers typical star ratings ranging from 2.8 to 4.3. Panel C in Figure 5.7 illustrates the magnitude of Marcello's adjustment—about two standard deviations on the IRT latent scale—and illustrates similar adjustments for five other restaurants with large IRT adjustments.[22]

Marcello's had closed by March 2025, but the Yelp website provides it an average score of 3.3 from thirty reviews. Despite the lack of information about Marcello's in the open dataset, bridging information from its one reviewer allowed us to do a decent job of estimating the restaurant's eventual score from a single review. Panel C in Figure 5.7 illustrates why. The average estimated threshold of Marcello's reviewer is among the lowest in the dataset: this reviewer also scored six other restaurants in the dataset, mostly large chains, and provided an average rating of 4.3 stars. This is enough information for the model to decide that this Yelper has low standards. As Panel B shows, the model considers this reviewer especially lenient, relative to the rest of the sample. Furthermore, it turns out that the point estimate for this user's upper threshold—the dividing line between his four- and five-star reviews—is 0.5, or roughly the second percentile of the distribution for that threshold parameter. Panel B shows that Marcello's story is quite general, in that we can predict a lot about how a reviewer will rate a specific restaurant if we know their level of average strictness. As the lowess curve illustrates, the extent to which the IRT adjusts restaurant scores is largely explained by average reviewer strictness.

It turns out the dataset contains a variety of different "types" of reviewers. We applied K-means clustering to the discrimination and threshold parameters estimated by the model—using the average estimated thresholds as a summary of leniency—and unearthed four types of Yelp raters: enthusiasts, softies, critics, and curmudgeons. Figure 5.8 illustrates these four groups. Enthusiasts and critics both discriminate reliably between different quality restaurants, but critics have higher standards than enthusiasts. Similarly, among raters who are not particularly reliable, we can differentiate between curmudgeons and softies, depending on whether they tend to indiscriminately heap love or hate on restaurants they visit.

[22]Panel C plots the three largest upwards and downwards standardized adjustments for restaurants that received at least ten reviews in our data (except Marcello's which received one review). The IRT estimate is standardized and the raw score is standardized to the IRT scale by first subtracting the raw mean and then dividing by the IRT estimate standard deviation. The IRT and raw standard deviations are similar: 0.62 versus 0.66.

Figure 5.8: K-means clustering of Yelp reviewers into four types based on leniency (mean threshold) and discrimination (IRT β parameter). *Panel A*: reviewers in leniency × discrimination space. *Panel B*: mean threshold profiles for each reviewer type; symbols mark the four threshold parameters (γ_1–γ_4) on the latent quality scale, with shaded bands indicating the corresponding star-rating region.

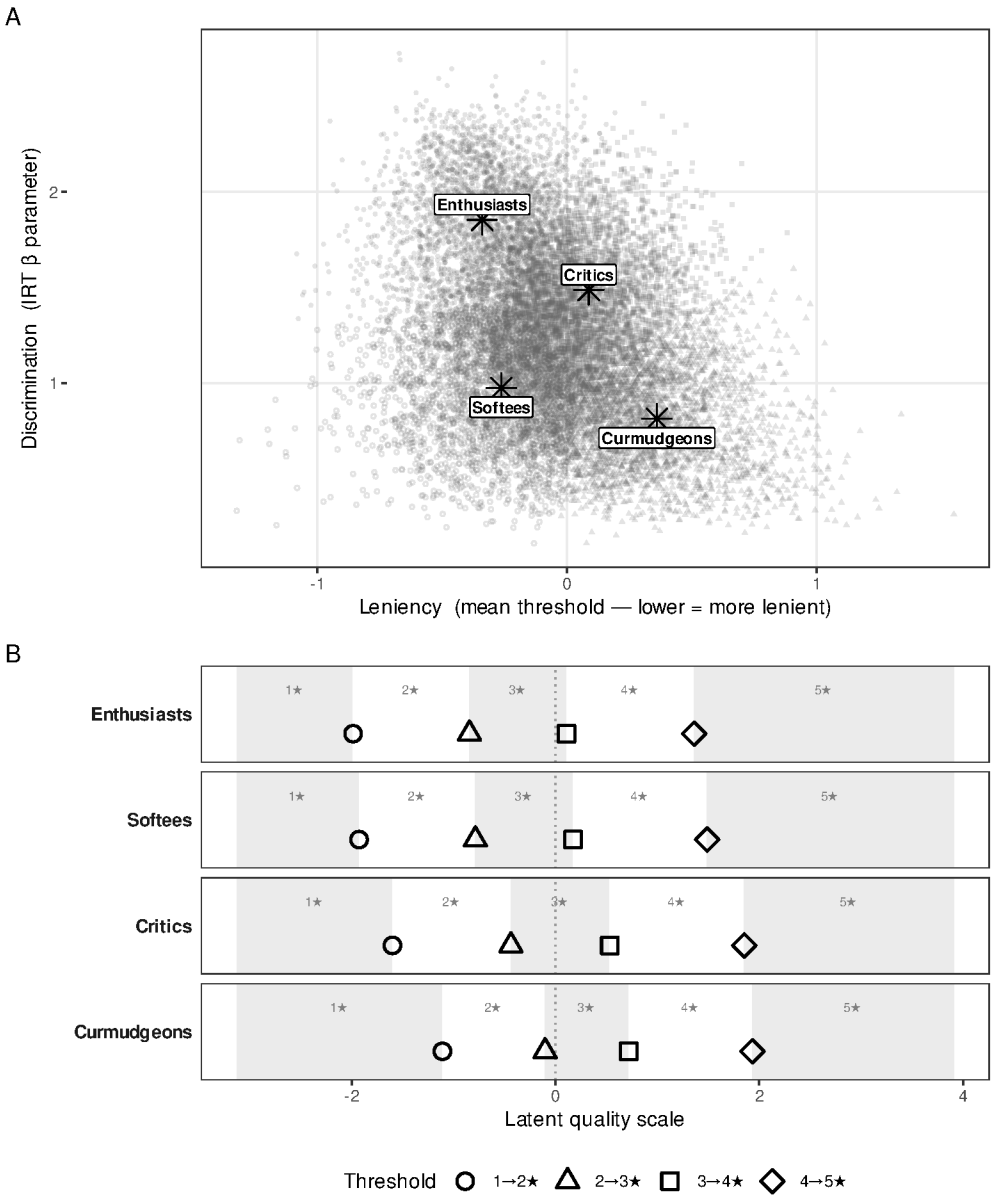
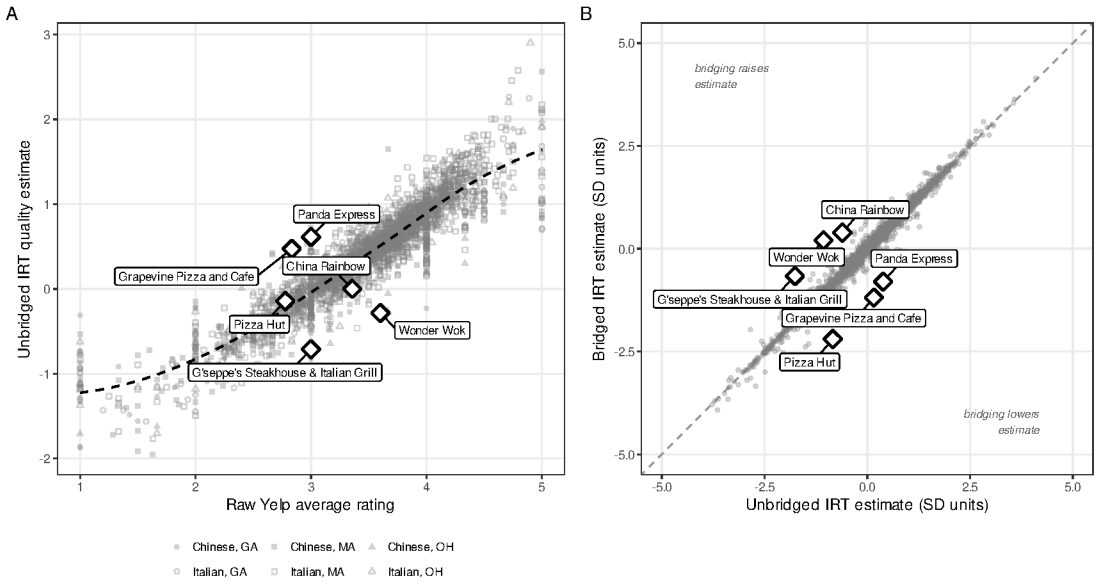


Figure 5.9: Yelp restaurant quality: unbridged IRT estimates and the effect of bridging. *Panel A*: Unbridged IRT latent quality estimates versus raw Yelp average ratings—the parallel to Panel A of Figure 5.7, but fit without cross-state bridge reviews. *Panel B*: Unbridged versus bridged IRT estimates, both expressed in the same standardized units; points above the 45-degree line are raised by bridging, points below are lowered. The six restaurants with the largest bridged-versus-unbridged shifts are highlighted with open diamonds in both panels.



So far, this is mostly just a story about DIF. Figure 5.9 illustrates how important bridging is to getting restaurant ratings right across states. Panel A in Figure 5.9 corresponds to the same panel in Figure 5.7, but now displays the results of fitting an IRT model to a subset of the data with all cross-state observations removed. Panel B compares the estimates that come out of this model to those from a model with cross-state bridges intact. We highlight the six restaurants whose relative position changes the most across the two models. In particular, we see that bridge observations produce big drops in the relative positions of big chain restaurants, while local haunts move up the rankings. This makes sense if we believe that big chains tend to be frequented by less discriminating raters while local haunts draw a more picky crowd.

Returning to Marcello's, the sole Yelper who rated this restaurant only assessed restaurants in the Atlanta area, so we might still worry that even the adjusted ratings are insufficiently bridged to tell someone from Boston, where standards for Italian food are likely higher than in Georgia, exactly where Marcello's stands. Of course, the pattern of bridging that we do have is quite complex. This user need not have scored any Boston restaurants for the data to be "sufficiently bridged." It could be that some of the other Yelpers who rated Atlanta area restaurants also evaluated Boston's offerings. Or they might have assessed other Atlanta area restaurants which were also rated by folks from Massachusetts. Knowing

whether a particular dataset is sufficiently bridged to validly compare two arbitrary units within the dataset is a complex question, lacking a simple answer. The rest of this chapter examines how much bridging we need and how to add bridges most efficiently for making valid comparisons between cases.

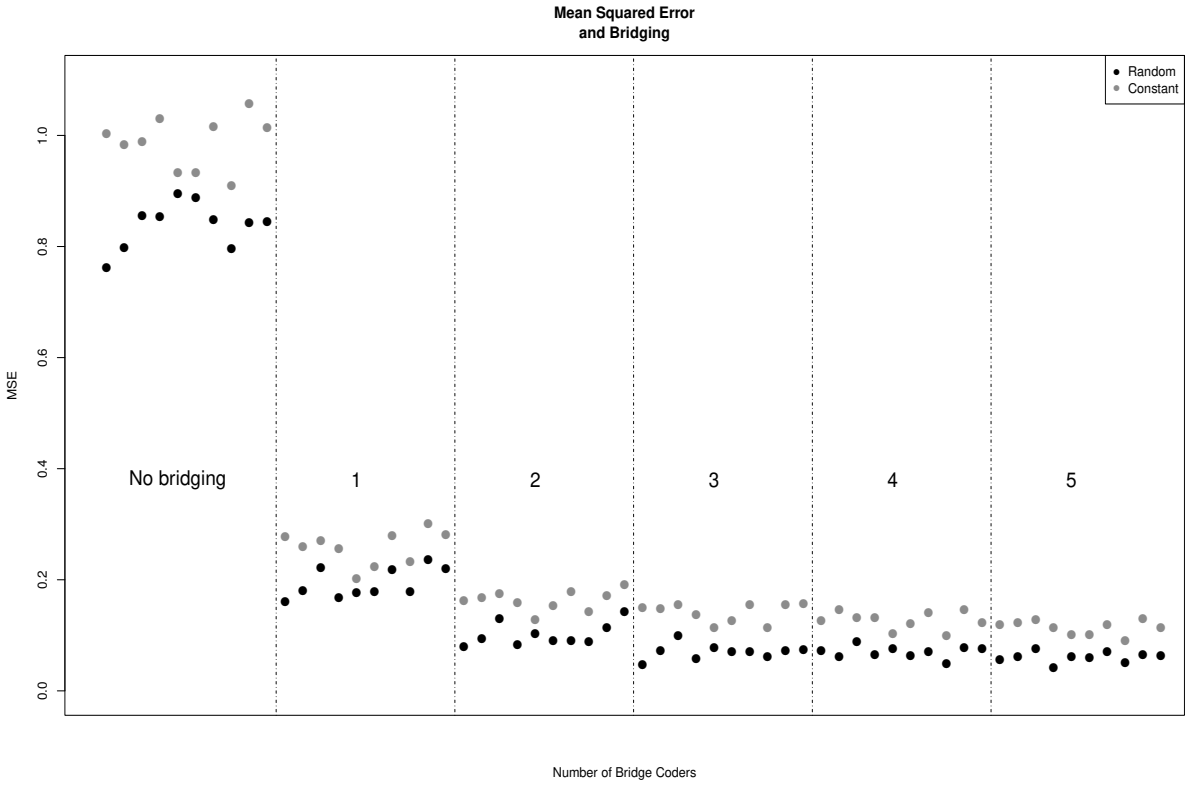
5.3 How To Bridge Efficiently

By this point, we hope you are convinced that bridging matters in many contexts and that overcoming bridging issues is crucial to effective measurement and inference from cross-case surveys. But it should also be clear that obtaining bridges can be difficult. Sometimes it is possible for respondents to provide "natural" bridges, as in the V-Dem or Yelp data, but the researcher often has little control over which bridges are available. In other contexts, like that facing the WHO team interested in understanding governance self-efficacy, tools like anchoring vignettes must be brought to bear because respondents cannot provide cross-case assessments. For other problems, such as the common-space ideal point estimation, one might need to be quite creative, or make strong assumptions, to find plausible bridges. Adding bridges will, in general, raise the degree of difficulty of fielding any given survey, and will often increase costs, both for the research team, and for respondents. Therefore, to most efficiently achieve cross-case comparability, we will often want to collect the smallest number of bridging observations necessary to achieve our goal.

So how much bridging do we need and what sorts of bridges provide the biggest bang for the buck? Pemstein, Tzelgov & Wang (2015) conducted a series of simulations to help better answer this question. This work was motivated by V-Dem, so focused especially on problems that confronted the project. In particular, it examines how much bridging one needs to get Likert scale equivalence for a single survey item, across two simulated countries when one has only a small number of respondents per country.

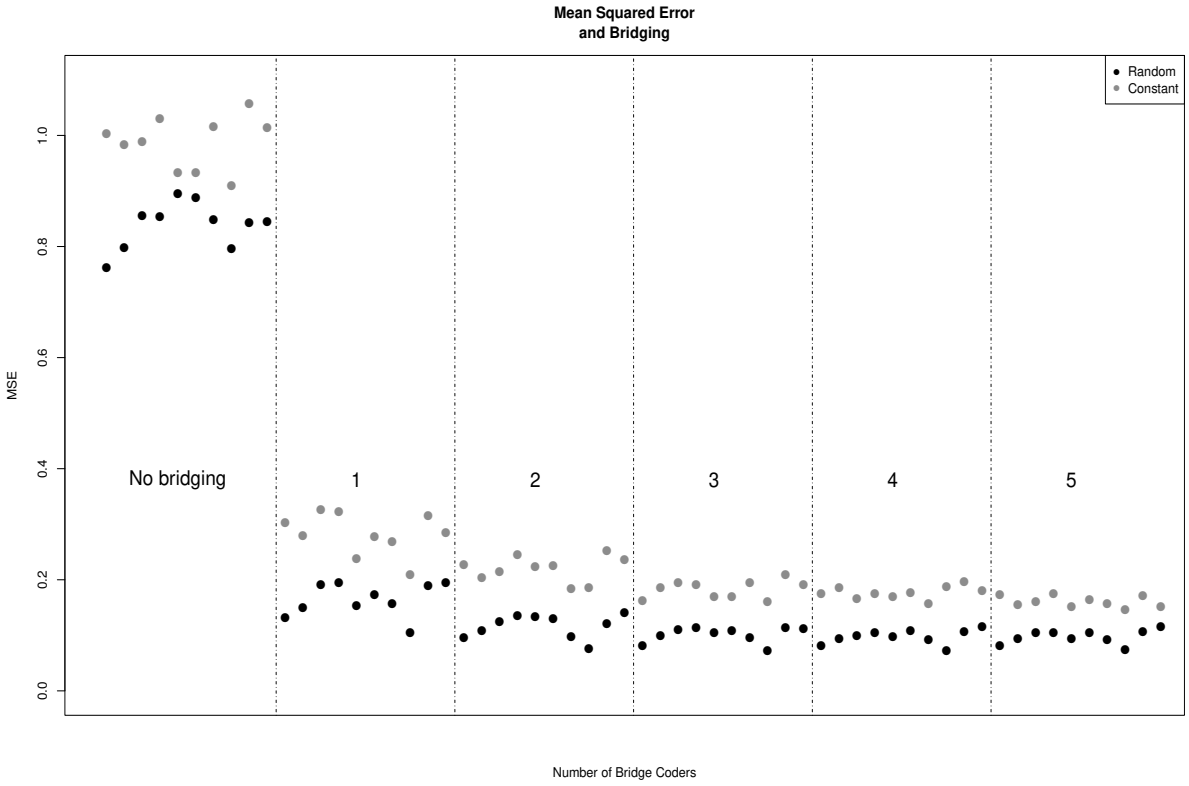
In one set of simulations, Pemstein, Tzelgov & Wang (2015) examine a particularly difficult version of the cross-country scale equivalence issue, something they dubbed the "Swiss" problem. This issue crops up when some cases exhibit relatively little variation, and therefore provide limited information about how individual respondents convert their perceptions of latent quantities into survey—in this case, Likert scale—responses. In V-Dem, this issue arises with countries, like Switzerland,[23] that have relatively stable institutions over long periods of time. "Swiss" respondents tend to never use large portions of the response scale, making it especially difficult for an IRT model to estimate their thresholds relative to those of respondents who use other parts, or all of the scale. Under these circumstances, prior assumptions about respondent thresholds dominate the data. And, because uncertainty about those thresholds is large, the prior for the latent trait also becomes especially important. This will tend to push latent variable estimates for "Swiss" cases toward the prior mean, regardless of the Likert score, as we saw in figure 5.4.

Figure 5.10: Evaluation of bridging from constant to random countries–mean square error of latent scores



Note: The figure shows the effects of bridging from a constant to a random country. Bridge coding is done for the entire time period (one hundred years). The vertical lines divide the figure into the bridging patterns, from a no-bridging to full-bridging (i.e., all respondents from the constant country also assess the random country). In each section of the figure, points represent the MSE for ten relevant simulations.

Figure 5.11: Evaluation of bridging from random to constant countries–mean square error of latent scores



Note: The figure shows the effects of bridging from a random to a constant country. Bridge coding is done for the entire time period (one hundred years). The vertical lines divide the figure into the bridging patterns, from a no-bridging to full-bridging (i.e., all respondents from the constant country also assess the random country). In each section of the figure, points represent the MSE for ten relevant simulations.

5.3.1 Thresholds drawn from the same distribution

Pemstein, Tzelgov & Wang's (2015) first set of experiments simulate true values on a latent scale for two hypothetical countries, with observations spanning 100 years for each simulated country. The first country is "constant," or "Swiss," with each true latent score simulated such that $z_{ct}^{constant} \sim \mathcal{N}(1.8, 0.05)$.[24] The second hypothetical country in each simulation exhibits highly random values for the latent trait: $z_{ct}^{random} \sim \mathcal{N}(0, 1)$. This initial set of experiments focuses on situations where respondent scale equivalence is relatively minimal.

Each simulation assumes that the respondents behave as an IRT model implies. Each respondent observes a given latent trait and then produces a 5-level Likert scale response based on thresholds generated from the following normal distribution: $\mathcal{N}(\{-1.5, -0.5, 0.5, 1.5\}, 0.2)$. In other words, the simulation assumes that raters had identically distributed, and quite even, thresholds. Note that *there is no bridging problem here* since respondent thresholds are drawn from a single distribution. Respondents are exchangeable, and simply taking normalized respondent means would provide an effective way to obtain scale-equivalent measures for each country-year. At the same time, a researcher cannot know that the two sets of respondents have similar thresholds ex ante. Without bridging, a traditional IRT approach (with a standard normal prior on latent traits) will under-perform naively taking respondents' assessments at face value, because of the tendency to pull "Swiss" cases toward the center of the distribution when the model relies on prior information to make up for limited (bridge) data.[25] But we have no easy way of knowing that we already have scale equivalence without obtaining bridges to help us estimate respondent thresholds.

Each simulation starts with no bridging—initially five respondents each assessed either one country or the other—and then proceeds to add bridges sequentially. Pemstein, Tzelgov & Wang (2015) examine how bridging affects the accuracy of latent trait estimates, coverage of 95 percent credible intervals for latent traits, and the accuracy of estimates of respondent's thresholds. Here, where the bridging problem is limited, they find that adding even one bridge drastically reduces the mean square error of the latent trait estimates, dropping by about 75 percent for the random country and roughly 80 percent for the constant country. As figures 5.10 and 5.11—which we replicate with the authors' permission—show, additional bridges produce only small additional gains. While we do not replicate the relevant figures here, coverage rates for 95 percent credible intervals behave similarly. While they start around 55 percent for random countries, and 75 percent for constant, they quickly jump up to 90 to 100 percent with a single bridge, regardless of whether the bridge comes from the constant country. Ideally, 95 percent coverage intervals should cover the true value around 95 percent of the time, so, under these simulated conditions, model performance approaches its maximum with the addition of only one or two bridges.

Interestingly, bridges from constant to random countries produce more bang for the

[23]Or Canada, in our previous example, and the discussion of the role of empirical priors in the V-Dem measurement model, in Chapter 4.

[24]Here the subscript c stands for country and t stands for time, or year.

[25]The V-Dem measurement model's empirical priors, which we discuss in Chapter 4, provide a way to moderate this tendency when bridge data are lacking. When information is limited they pull estimates toward what one would achieve by taking confidence-weighted response averages, rather than toward the center of the distribution. But as more data becomes available—especially bridge observations that allow the model to better estimate respondent thresholds on the global scale—these data can overwhelm the prior.

buck than bridges going the other way. It is more valuable to have a few constant-country respondents assess the random country than to have random-country respondents assess the constant country. This happens because constant-to-random bridging provides substantially more information about respondents' threshold parameters than does its counterpart. Of course, this makes perfect sense. Without bridges, we know very little about constant-country respondents' thresholds because their survey responses exhibit little variation. We learn a lot more by having those respondents assess a case with a lot of variation, forcing them to use more of the Likert scale, than we do by having a well-estimated respondent come over and assess the constant case. This result illustrates a key rule of thumb for adding bridges to your survey data. Whenever possible, have respondents—especially those who assess cases with little variation—bridge cases with substantial variation. A corollary point is that it can be profitable to elicit variable survey responses from respondents who do not naturally provide much variation, although the value of adding bridges to maximize variation for a given respondent depends on how much information we have about the cases that they already assess.

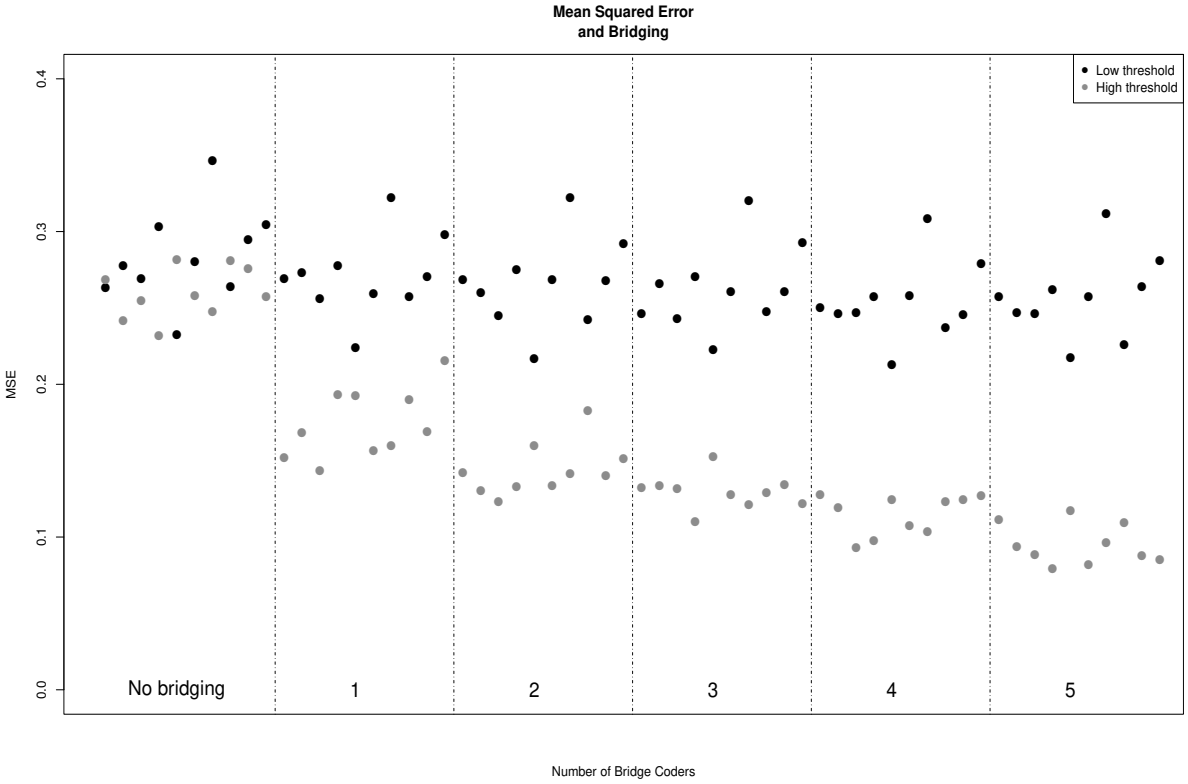
With real bridges, maximizing response variation can pose a challenge, since it may not be possible to meet this goal given respondents' capabilities and/or variance in the available cases. For example, our Acapulco-Cancún example illustrates a situation where we have little case-level variance to exploit. Anchoring vignettes can shine here, because they provide a tool that, when well-designed, allows one to observe your respondents' assessments of cases that span the underlying latent scale.

5.3.2 Thresholds drawn from different distributions

Pemstein, Tzelgov & Wang's (2015) second set of experiments focus on the bridging problem, but in a context where cases (here, countries) exhibit substantial over-time variation. In these experiments, the simulations again involve two random countries, as defined above. Here, low-threshold respondents—whose thresholds are drawn, ordered, from a $\mathcal{N}(-1, 0.5)$ distribution—evaluate the first country. These respondents can discriminate the bottom end of the scale, but not the top. They are, in other words, quite lax. The respondents for the second country, on the other hand, are rather strict. They have thresholds drawn from a $\mathcal{N}(1, .5)$ distribution—again subject to an ordering constraint.

Figure 5.12, from Pemstein, Tzelgov & Wang (2015), shows what happens when the simulations incrementally add bridges from one country to the other, in this case from "low" to "high." We see that the mean squared error of the country receiving bridges drops gradually. On the other hand, these bridges tell us little about the country that receives no bridge assessments. When respondents only use parts of the scale, and therefore simply cannot discriminate certain levels of the latent trait, it is difficult to produce scale-equivalent assessments of cases without ensuring that different types of raters assess each case. The simulations presented here represent an extreme version of this problem, much like the bifurcation in glass half-full and half-empty respondents in our Acapulco-Cancún example, but they illustrate a general guideline. We can sometimes maximize efficiency by carefully obtaining a small subset of bridges. But if we suspect that the respondents that assess a particular case are particularly optimistic or pessimistic, and therefore probably cannot discriminate across the entire latent scale, then it is crucial to have a diverse set of bridge

Figure 5.12: Evaluation of bridging from a country with low respondents' thresholds to a country with high respondents' thresholds, mean square error of latent scores



respondents assess that case.

5.3.3 Automating bridge selection

The simulations in the previous sections illustrate general rules of thumb for bridge selection. Broadly speaking, researchers should strive for exceptionally dense bridging, since it is the strategy most likely to yield accurate and reliable results. But, in projects like V-Dem, where surveys demand significant investment from experts, we want to use patterns in existing ratings that we already have to identify potential bridges that will likely give us the most bang for the buck. Our approach to this problem is motivated by the scholarship on computerized adaptive testing (CAT).[26]

In the testing context, CAT techniques seek to present questions to a student in an order that identifies that student's latent ability on a particular dimension as quickly as possible. This shortens test time (or survey length) and therefore conserves resources. In this scenario, students are analogous to country-years and test questions take the place of respondents. The researcher generally assumes that question (respondent) parameters—each $\gamma_{r,k}$ and β_r—are known[27] and iteratively chooses questions based on estimated parameters that minimize the expected posterior variance in the estimate of the student's latent ability. At the beginning of an exam process, CAT algorithms prioritize highly discriminating questions. Then questions are selected based on the relationship between item thresholds and the current estimate of the student's latent ability. Intuitively, this approach generally results in the algorithm presenting a student with a moderately difficult question to begin with, presenting a harder question if the student does well on the first question, or an easier question if the student does poorly on the first question, and then repeats the process until the estimate of the student's latent ability reaches some desired level of precision. A common strategy for item selection uses the MEPV (minimum expected posterior variance) criterion. Specifically, for a single student with latent trait, z, at each step in the process, the algorithm chooses the question, q, that meets the condition

$$\underset{q}{\operatorname{argmin}} \left[\sum_{k=1}^{K} p_q(y_q^{i+1} = k|\mathbf{y}^i) \mathrm{Var}(z|\mathbf{y}^i, y_q^{i+1} = k) \right] : q \in \mathbf{Q}_i \tag{5.4}$$

where $p_q(y_q^{i+1} = k|\mathbf{y}^i)$ is the posterior predicted probability, prior to answering question q, that the student's answer to question q achieves a score of k, $\mathrm{Var}(z|\mathbf{y}^i, y_q^{i+1} = k)$ is the posterior variance of the student's latent ability, conditional on her previous performance and achieving a score of k on question q, and $\mathbf{Q}_i$ is the set of remaining potential questions at the ith step of the process (Choi & Swartz 2009).

[26] See Montgomery & Cutler (2013) for an introduction to this scholarship aimed at survey researchers in political science. Choi & Swartz (2009) provide an overview of CAT item selection methods for ordinal items.

[27] In this context, the researcher has fit an IRT model to oodles of testing data from previous test-takers and has precise estimates of these parameters for the set of available questions.

Computerized adaptive bridging

Our problem differs from the traditional CAT setting in two key ways. First, while CAT focuses on the latent trait for a single student, we are trying to improve all of our latent trait estimates by making them comparable. This means that we are trying to select combinations of cases and respondents—that is, bridges—that we expect to tell us as much as possible about latent trait parameters as a whole. Indeed, a well-chosen bridge could substantially shift numerous latent trait estimates at once, as the implications of that bridge cascade through model parameter estimates, as Figure 5.10 and Figure 5.11 demonstrate.

Second, CAT can rely on a well-calibrated IRT model when one trusts that the item (respondent) parameters—discrimination and thresholds—are both unbiased and precisely estimated. But we cannot trust those parameters in an insufficiently bridged model; the bridging problem exists precisely because we lack the information necessary to estimate those parameters on a common scale. Nonetheless, CAT can help us to make good guesses about which missing observations will most efficiently bridge our data. When a bridging problem exists, good bridges will substantially alter parameter estimates, particularly z_{ct} and $\gamma_{r,k}$ values. Good bridges are cases where a bridge respondent behaves in a way that is highly inconsistent with the model's expectations. In a sense, we are looking for bridges that have the potential to *surprise* us, given the current model.

We consider one greedy algorithm[28] that works by iteratively seeking "surprising" bridges. The MELPD—or maximum expected local probability difference—algorithm iteratively chooses bridges that have the highest potential for local positive change in posterior predicted probability. At each step, MELPD selects the bridge that meets the condition

$$\underset{\{c,t,r\}}{\operatorname{argmax}} \left(\sum_{k=1}^{K} \left[p_{ctr}(y_{ctr}^{i+1} = k|\boldsymbol{\theta}^i) \left[p_{ctr}({y_{ctr}}^{i+1} = k|\boldsymbol{\theta}^{i+1}) - p_{ctr}({y_{ctr}}^{i+1} = k|\boldsymbol{\theta}^i) \right] \right] \right) : \{c,t,r\} \in \mathbf{B}_i \tag{5.5}$$

where $\boldsymbol{\theta}^i$ is the vector of fitted model parameters given $\mathbf{y}^i$ and

$$p(y_{ctr} = k|\boldsymbol{\gamma}, \mathbf{z}, \boldsymbol{\beta}) = \Phi(\gamma_{r,k} - z_{ct}\beta_r) - \Phi(\gamma_{r,k-1} - z_{ct}\beta_r) \tag{5.6}$$

is the posterior prediction, from a fitted model, for case $\{c,t,r\}$. Intuitively, this algorithm looks for bridges that elicit ratings that make a lot more sense after the fact than they do before they are observed. This is a potentially fruitful approach because, absent bridging, the model may imply that certain potential ratings are very unlikely, but, after observing such a rating, may judge it quite unremarkable. For example, consider a potential bridge which involves a respondent from a high-threshold country rating a case that respondents in a low-threshold country have routinely judged to be at the top of the scale. Absent bridging, the model would reasonably predict that the high-threshold respondent would also assess that observation highly. But a low rating by the bridge respondent should encourage a large

[28] A greedy algorithm attempts to find a globally optimal solution to a given problem by iteratively making locally optimal choices. In this context, this means choosing individual bridges that maximize (or minimize) some criterion at each step. In other words, these algorithms do not look multiple steps ahead. We focus on greedy algorithms here, both because of their computational tractability, and because they match a process of iterative bridge recruitment that is likely to be easiest to implement in a real study.

shift in threshold estimates, and this previously unlikely rating would look quite reasonable in the context of the re-calibrated model.

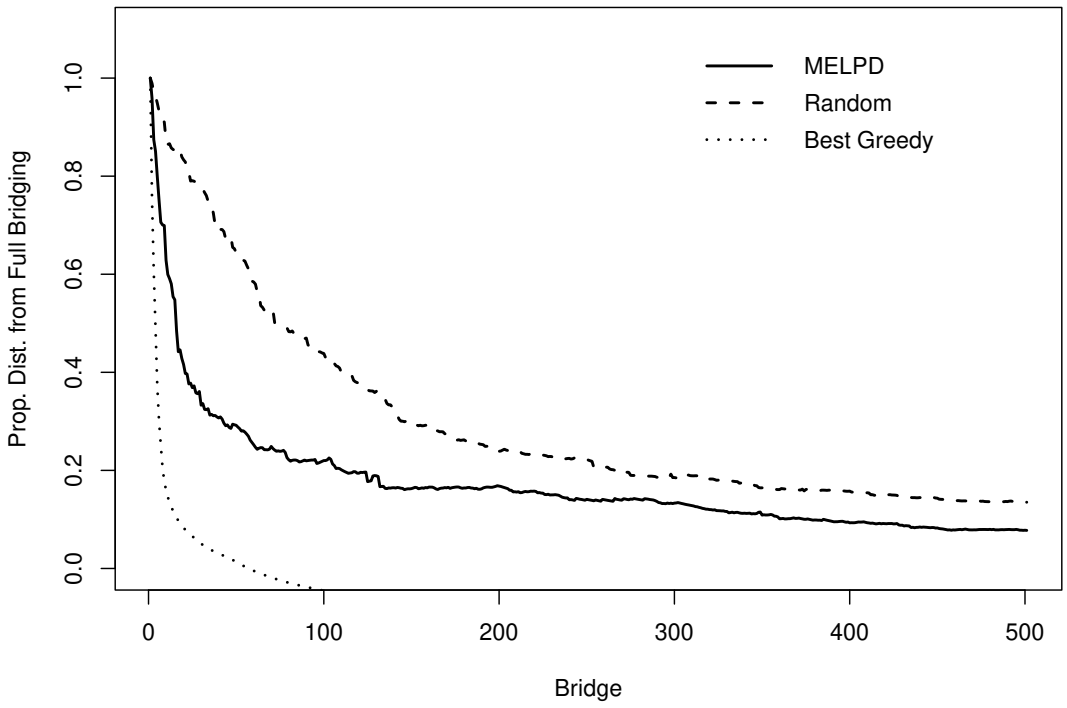
Note that we continue to multiply these potential "surprises" by the ex ante predicted probability of each potential rating at each step. Specifically, we continue to use our current model to produce best guesses of what each bridge observation will look like, prior to eliciting that observation. We experimented with a version of this algorithm that excludes such weights, so as not to prioritize "surprises" that seem most likely to a given—probably badly calibrated—model. While the two algorithms perform similarly to one another when there are few bridge observations, including these best guesses improves performance once bridging become relatively more dense. That is, this weighting scheme does not seem to hurt us too badly when the model is especially wrong and pays dividends as the model becomes better. We also experimented with a global version of this algorithm—one that looks at shifts in all predicted probabilities instead of focusing on the local effect of any given bridge observation. Intuitively, such an algorithm might capture the cascading effects of specific bridges. But somewhat surprisingly, we found our locally focused algorithm outperformed its global counterpart. Nonetheless, future research might look at more restricted generalizations of MELPD. For instance, a researcher might prioritize bridge observations that maximally change posterior predictions for all respondents on both sides of the bridge.

In principle, we should be able to run these algorithms until the observed mean squared error (MSE) between parameters from the pre- and post-bridge model fits falls below a desired threshold. Moreover, we can easily adapt this algorithm for constrained search, by altering the contents of $\mathbf{B}_i$. For instance, while an expert on parliamentary politics in Ecuador might, in principle, provide a very informative bridge were she to assess legislative corruption in Bangladesh in 1964, we are unlikely to find an Ecuadorian specialist with detailed knowledge of Bangladeshi politics in the middle of the last century. Similarly, it will be impractical to recruit bridges serially in several different contexts. So rather than attempting to recruit the single bridge that maximizes Equation 5.5, we could use this approach to identify the set of n most potentially informative bridges, from a set of plausible potential bridges, at each step.

The MELPD algorithm is computationally expensive because it requires us to generate posterior estimates for the model parameters for each potential outcome of each potential bridging rating at each step of the search. These computational needs quickly become prohibitively expensive when we use Markov chain Monte Carlo (MCMC) techniques—even ones that generally converge quickly, like Hamiltonian Monte Carlo (HMC)—to estimate model parameters. Standard CAT applications only require the estimation of a single parameter, making numerical integration a practical approach (Montgomery & Cutler 2013); here we need to estimate full multi-parameter models, making this solution a non-starter. Therefore, we use a variational approximation algorithm to estimate model parameters at each step (Jordan, Ghahramani, Jaakkola & Saul 1999).[29] Appendix 5.A provides details.

[29]Computer scientists developed variational approximation techniques, and their descriptions of these methods tend to use jargon that is unfamiliar to social scientists. Ormerod & Wand (2010) provide an overview of these methods aimed at statisticians, using language that should be familiar to social scientists with statistical training. Grimmer (2011) provides an introduction using examples drawn from political science.

Figure 5.13: Comparing CAT-inspired bridge selection algorithm



Evaluating the MELPD algorithm

Figure 5.13 depicts the results of a Monte Carlo experiment that compares the performance of the MELPD algorithm to random bridge selection and a best-possible greedy algorithm. We simulated twenty four-country datasets using a procedure similar to what we did when examining the constant response problem in Section 5.3.1. Here, we have two random countries and two constant ones, and we simulate fifty years of data for each country. Consequently, the unbridged dataset includes 750 potential bridges, as each of the five respondents for each of the four countries can assess 150 country-years in the other three countries.

We simulated three search processes. Each process sequentially added five hundred bridges[30] to the unbridged dataset. After each bridge was added, we calculated the mean square error (MSE) between the true distribution of latent traits and the estimates provided by the ordinal IRT model. We then computed the proportional improvement in MSE by dividing the difference in MSE between the current model and the fully bridged model by

[30]Here, each bridge is a single country-year observation, rather than a full 100-year time series, as in Section 5.3.1. In this respect it resembles the "lateral coding" discussed earlier, in which an expert rates a single year for many additional countries; standard V-Dem bridge coding, by contrast, typically spans longer periods.

the difference in MSE between a completely unbridged model and the fully bridged model. The lines on the plot represent median performance across simulations.

In the first search process, we chose bridges at random (the long-dashed line). In the second, we used the MELPD algorithm to sequentially select bridges that had the maximum average potential to alter latent trait estimates (the solid line). Finally, the best-possible algorithm always chose the bridge that minimized the MSE between true and estimated latent traits, at each bridging step, to help us visualize the bound on achievable greedy-algorithm error reduction (short-dashed line).

Figure 5.13 first shows that it is possible (in principle) to eliminate bridge problems with a relatively small number of carefully chosen bridges. The best greedy algorithm eliminates over 90 percent of the bridging error with only twenty bridges. By one hundred bridges, the error is gone. This result is leaps and bounds better than what happens with the random algorithm, which has only eliminated about 20 percent of the bridging error at twenty bridges, and about 55 percent at one hundred bridges. The MELPD algorithm starts out strong, closely tracking the best greedy algorithm for the first few bridges and reducing bridging error by roughly 60 percent with twenty bridges. At one hundred bridges, it has eliminated about 80 percent of the error. Then performance flattens out, and the random algorithm slowly closes the gap over the next four hundred bridges. At this point, most differences in performance between the MELPD and best greedy algorithm are likely driven by idiosyncratic aspects of the data, rather than bridging issues. Indeed, actual MSE is generally around 0.2 for the MELPD algorithm by one hundred bridges, which is only about five percent of the normal scale. Furthermore, examinations of scatter-plots of fit show that much of the additional improvement of the best greedy algorithm is in the random countries, which are not poorly calibrated from a bridging standpoint. In sum, MELPD likely provides a powerful tool for efficient bridge selection when full, or even dense, bridging is infeasible. More broadly, we present it both as a practical algorithm that researchers can deploy and as a way to reason about *which* bridges are most informative—and thus as a benchmark for evaluating the efficiency of any bridge-selection strategy.

5.4 Identifying Bridging Problems in Real Data

The simulations in the previous section starkly illustrate the scope of the bridging problem and provide some intuitions about how to select bridges to mitigate this issue. The experiments with simulated data suggest that bridge ratings help solve the issues of "constant" cases and different thresholds of different groups of respondents. When the quantity of bridge coding increases, the MSE between the country-year estimates and the simulated latent scores decreases. But bridging's ability to improve estimates is contingent on the scope of the problem. When DIF is minimal—when respondents use the survey item response scale similarly—then we can rapidly achieve scale equivalence with just a few bridges. But when respondents exhibit substantial cross-case DIF, our simulations demonstrated that bridges must flow in both directions to identify respondent thresholds and achieve cross-case scale equivalence for all cases; unidirectional bridging tends to improve estimates for only some of them. Moreover, case characteristics matter. When some cases exhibit little variation, it can be especially helpful to obtain bridging observations for those cases' respondents, since

we learn very little about their thresholds from their ratings of their primary case. This is somewhat counter-intuitive, because we know more about thresholds of respondents who focus on high-variation cases; in theory they might provide especially informative bridges. But it turns out that—at least with time-series cross-sectional data formats like V-Dem—*it may be best to focus on filling in the model's understanding of rater thresholds even when one is most focused on latent trait assessment.*

In any case, our ability to evaluate bridging quality in these simulations rests on the fact that we know true latent values. To understand whether we face a bridging problem in real-world applications, we need techniques for identifying poorly bridged cases from observable quantities. So how can one know whether a particular dataset has a bridging problem? We suggest three approaches:

1. Evaluating bridge density.
2. Assessing face validity across cases.
3. Conducting posterior predictive tests.

5.4.1 Evaluating bridge density

First, ask yourself how much cross-case bridging you have. How do respondents cluster across cases? Often clustering will be quite distinct, and the potential for bridging issues will be evident. The WHO survey data on political efficacy, for example, are clustered perfectly by country without vignettes. Congressional and Supreme Court voting records are distinct in the absence of clever bridging strategies, such as coding legislators' positions in amicus briefs. In other cases, such as V-Dem surveys, respondents will cluster substantially, but some respondents will assess multiple cases, providing some bridging. Sometimes, you will be lucky and most of your data will be densely or even completely bridged (something would be completely bridged when every respondent evaluates every case). When bridging is incomplete, ask if any case is altogether unconnected from the broader bridging network. That is, are any cases evaluated by a set of respondents who evaluate no other cases? Under such circumstances, the model will have to rely purely on prior information to identify respondent thresholds, and latent variable estimates for such "islands" are likely to be incorrectly scaled relative to the broader set of cases. If no isolated cases exist, next look for "continents." Taking the bridging metaphor a touch literally, ask if one could walk across bridges from any case to any other case in the dataset? If not, then you will have one or more continents in your data and sets of cases will be locally bridged, but disconnected from the broader case network. These case-continents are likely to exhibit local scale identification, but scales will likely differ across continents.

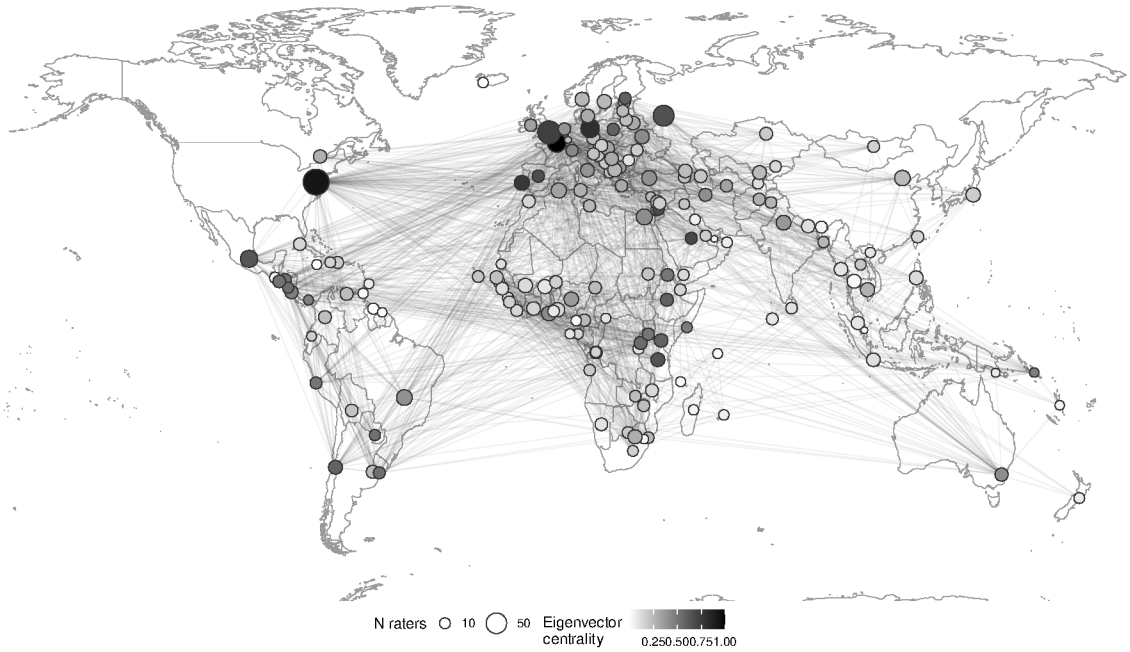
Figure 5.14 illustrates island and continent finding, in R, with the response matrix for V-Dem's "Freedom from political killings" variable, and with anchoring vignettes removed. Each row in this matrix is a country-regime,[31] while each column represents a respondent. Matrix elements represent the assessments that raters provided for each country-regime.

[31] See Chapter 7, which explains how V-Dem collapses country-date level data into "regimes." All that matters here is that these regimes represent spans in time, and each row of the data matrix represents a single country-timespan observation.

Figure 5.14: R code for finding islands and continents in response matrices

```
#### Load ratings matrix, without vignettes, country-obs on rows, respondents on columns
novign <- readRDS("v2clkill_novign.rds")

#### First create a list of respondents (column #s) for each country
#### We will assume countries have sufficient within-country overlap to be
#### bridged over time, so we can collapse to the country level.
countries <- sapply(strsplit(rownames(novign), "-"), function (x) x[1])
country.resp <- sapply(unique(countries), function (x) {
 tmp <- novign[x==countries,] ### all the rows associated with the given country
 if (is.null(dim(tmp))) ### one-row country
 which(!(is.na(tmp)))
 else
 which(apply(tmp, 2, function (x) any(!is.na(x))))
})

#### Create an adjacency matrix from the ratings matrix.

### Empty matrix
amat <- matrix(0, nrow=length(unique(countries)),
 ncol=length(unique(countries)))
colnames(amat) <- names(country.resp)
rownames(amat) <- names(country.resp)

### Go through the country/respondent list and fill in adjacency matrix
for (i in 1:nrow(amat))
 for (j in 1:ncol(amat))
 if (length(intersect(country.resp[[i]], country.resp[[j]])) > 0)
 amat[i,j] <- 1

#### Islands have no edges with any nodes but themselves
which(apply(amat, 1, sum) == 1)

#### Find continents by finding all connected subgraphs

### Load graph library
library(igraph)

### Create a graph from the adjacency matrix
graph <- graph_from_adjacency_matrix(amat, "undirected")

### Find continents (connected subgraphs/components)
continents <- components(graph)

### Print the sub-graphs
groups(continents)
```

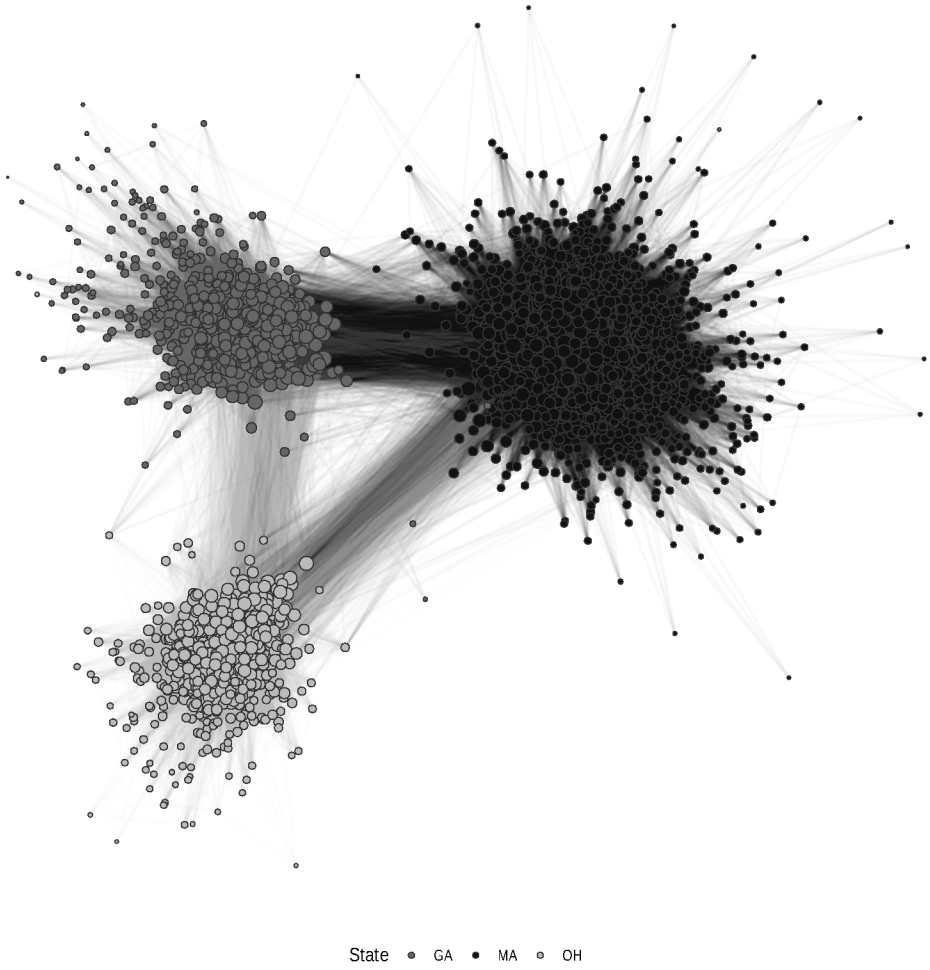
Figure 5.15: Bridge network for V-Dem's "Freedom from political killings" variable. Each node is a country, placed at its capital city; node size reflects the number of expert raters who assessed that country, and node shading reflects eigenvector centrality (darker = more central). Edges connect pairs of countries that share at least one rater.



These data are quite sparse, and include many missing values, because most respondents do not evaluate most countries. The code in Figure 5.14 identifies continents and islands by first converting the assessment matrix into a graph where countries serve as nodes, and respondents who assess multiple countries are edges. Specifically, when any respondent r provides an assessment of both countries i and j, we say that there is an edge between i and j. If no respondent assesses both i and j, then there is no edge between these two nodes. The code works by first converting the assessment matrix to an *adjacency matrix*. This is a square matrix with nodes—that is, countries—identifying both the rows and columns. Each element (i, j) of this adjacency matrix equals 0 if there is no edge between i and j, and 1 if there is an edge between the two nodes. We then use the `igraph` library to convert this matrix into a graph representation, and use the `components` function to find every connected subgraph within this graph. In other words, it finds every subset of nodes which is connected to every other node—perhaps indirectly—by at least one edge. In this context, it finds every subset of countries that is connnected to every other country in the subset, by at least one bridge assessment.

In the case of the "Freedom from political killings" data, it turns out that there is one large continent and four islands. That is, every country in the world—except for Bahrain,

Figure 5.16: Bridge network for Yelp restaurants in Massachusetts, Georgia, and Ohio. Each node represents a restaurant, colored by state. Edges connect pairs of restaurants reviewed by at least one common Yelper, colored by that reviewer's home state (the state where they wrote the most reviews). Node size reflects the square root of the restaurant's degree (number of shared-reviewer connections). The three state clusters are clearly visible, with a smaller number of cross-state bridges connecting them.



East Germany, Hong Kong, and Iceland—is connected by a network of bridge assessments. These four countries are rated only by respondents who assess no other countries. Figure 5.15 provides a visualization of the bridging network for this variable. Each country is represented by a circle at its capital's location on the map. The size of the circle represents the total number of experts who evaluate the given country, for any year. Each country is shaded according to its eigenvector centrality, which quantifies the extent to which any given node is connected to other nodes, while weighing connections to highly connected nodes more than connections to nodes with fewer connections. This is one popular way to represent how important—or "central"—a node is to a given network. Here, we see that France, the US, Germany, Portugal and the UK play especially pivotal roles in our bridge network, which makes sense given their linguistic, socio-political, and educational roles in the global system. But clearly other countries, such as Russia, Saudi Arabia, Mexico, Tanzania, and India are crucial to bridging V-Dem's expert network within sub-regions. Note that V-Dem relies on anchoring vignettes to supplement this bridge network, both linking the islands to the main continent, and adding substantial density to the bridge data. This is crucial because, while many cases exhibit dense bridging, others—in addition to the un-bridged islands, Guyana, Jamaica, Paraguay, and Suriname are examples of weakly bridged cases—have few natural bridges. As we will see in the next chapter, it also helps to ensure that informative bridges, spanning the latent variable space, exist.

Figure 5.16 provides a visualization of Yelp's bridging network, where we operationalize the bridging problem in terms of systematic differences in respondents from the three states (Georgia, Massachusetts, and Ohio) where we have restaurant data. Standards for Chinese and Italian restaurants may look quite different in Massachusetts than in the other two states, for instance. We can see three large, densely connected sets of restaurants in each state. Some restaurants in each state are only loosely connected to the broader network. These are generally restaurants with few overall ratings. While connections between cross-state edges are somewhat less dense than within-state connections, there is still substantial cross-state bridging. The shading in the figure shows that Georgia Yelpers travel less than their counterparts in Massachusetts and Ohio, but at least some reviewers from all three states reviewed restaurants in the other two states. The edged networks between the clusters are dominated by Massachusetts and Ohio raters, with relatively fewer Georgia respondents rating outside of their state.

Generally speaking, bridge network quality is complex and relies on multiple factors beyond whether all cases are connected. Table 5.2 summarizes several bridge network metrics for the Yelp and V-Dem "Freedom from political killings" datasets. "Node degree" is the number of cases to which each case is directly bridged, while "graph density" measures the fraction of all possible case-pair edges that are present in the bridge network, essentially normalizing node degree by the number of cases. "Bridge rater fraction" is the proportion of raters who assess more than one case, and "mean shared raters per edge" counts how many raters, on average, connect each bridged pair of cases. By most measures, the Yelp data are more densely bridged at the rater level, although V-Dem has higher raw graph density. Every Yelp reviewer assesses multiple restaurants and edges are denser—that is, bridges are replicated by more respondents. V-Dem's lower bridge rater fraction (32 percent) and thinner edges reflect the specialist structure of expert surveys, where most coders assess only their primary country. V-Dem specifically worked to build a bridging network that

Table 5.2: Bridge network density metrics for the Yelp and V-Dem datasets.

	Yelp	V-Dem
Cases	3,093	182
Raters	13,615	1,445
Mean node degree (SD)	292.5 (247.3)	24.2 (16.8)
Graph density	0.095	0.134
Bridge rater fraction	1.000	0.316
Mean shared raters/edge	2.33	1.76
(normalized by raters/cases)	0.53	0.22

would cover most cases by providing its experts incentives to provide bridge ratings, and this bridging is reflected in the denser graph. But only some experts have the capacity to rate multiple cases, concentrating bridges among a subset of raters. Bridge rating is more natural for respondents in the Yelp dataset and emerges organically in that context.

When evaluating bridge density, note that certain bridges are more valuable than others. For instance, consider V-Dem's "Freedom from political killings" variable and two hypothetical countries: Country A has five experts who gave the country the highest score (of five) across the time period, while Country B has five experts who all gave it zeros for the entire period. Say one respondent from each country also assessed the other country, but the countries have no other bridges. Here we would have a decent amount of bridging, but we would lack information about respondent thresholds in the middle of the scale. Thus, while we will likely get the rank order of the two countries right, their overall placement on the latent dimension will likely be too close to the center of the distribution, as we lack the information necessary to place thresholds across the broader scale. Generally speaking, asking respondents who assess such "constant" cases to respond to questions about highly variable cases can yield substantial dividends (as we saw in the previous sections' simulations). Thus, it can be important for a researcher to go beyond evaluating their number of bridges and which cases the bridges connect, and to also consider what sorts of bridges are available. All else being equal, bridging observations that turn respondents for whom the model has only vague information about thresholds into respondents for whom the model can tightly estimate thresholds will yield higher dividends.

5.4.2 Assessing cross-case face validity

When V-Dem collected its pilot data, it included no bridges in the data collection plan and experts assessed only one case. After this initial data collection phase, the V-Dem team held a conference where project and regional managers[32] looked at simple averages and standard deviations of expert responses to evaluate initial data quality. A theme quickly emerged: country time series looked quite reasonable and historically accurate but any time

[32]V-Dem project managers oversaw survey development while regional managers provided regional expertise and helped to identify and recruit country experts.

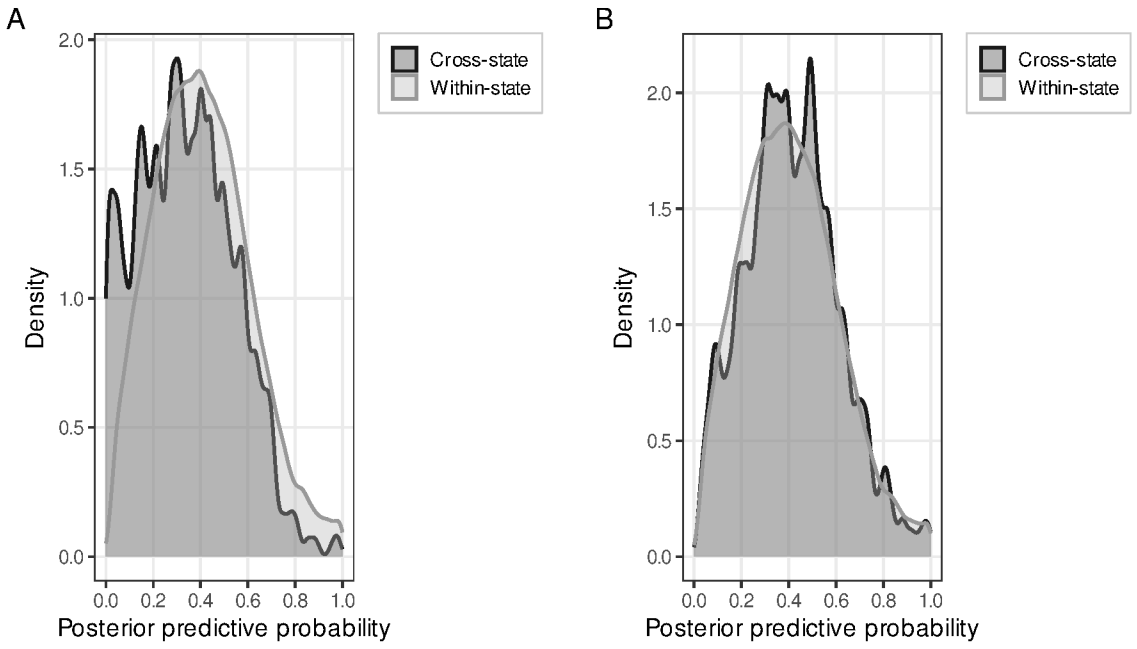
the estimates for two disparate countries were put on a screen, at the same time, confusion often resulted. In particular non-democratic countries and emerging democracies (especially countries exhibiting substantial variation in over-time ratings) often showed periods where their average ratings for a given variable outstripped ratings for established democracies. In one example, recent media freedom in Pakistan was judged higher than media freedom in the United Kingdom. Examining individual ratings, it quickly became clear that the Pakistani and British experts were operating from different baselines. Indeed, the scaling issue that we illustrated above with Canada and Germany—and again with the random-constant country simulations—emerged again and again in the V-Dem dataset. It was obvious that surveyed experts understood V-Dem's questions differently across cases, and that V-Dem's response scales lacked anchoring. This lack of face validity highlighted the importance of bridging and anchoring to the long-term success of the V-Dem project. Presenting these data helped V-Dem's methods group convince the larger project team of the value of inviting respondents to provide bridge ratings and, eventually, anchoring vignettes.

This anecdote illustrates the value of cross-case validity checks for revealing bridging problems. Such checks are perhaps easiest in the panel data world, where one can look at both the within- and between-case pattern of estimates. In the V-Dem case, the fact that within-country series looked quite reasonable, but relative series placement looked wrong, was readily identifiable. But, in general, one may have reasonable a priori expectations about the rank orderings of at least a subset of cases. If such expected rank-orderings are not reflected in latent trait estimates, then one has prima facie evidence for scale incoherence, and the likely cause of such scale incoherence is a lack of sufficient bridging. We therefore suggest checking the rank order of estimates against informed expectations. This approach may not be possible in all contexts. For example, our imagined Acapulco-Cancún vacation comparison lacked reference cases that we could use to identify face validity issues. But if we had added Newark, New Jersey—a place where virtually no one goes on vacation—to the mix and found it fared better than Acapulco, we might have realized that something was awry without knowing the underlying data-generating process. There is no one-size-fits-all process for assessing the face validity of one's estimates. That said, it often makes sense to include a "Newark" or two—that is, cases likely to diverge from your target cases in their level of the latent concept (here, enjoyment as a tourist)—in your data collection process, even if such cases are not your core object of interest. While one generally conducts measurement exercises to learn something new about the world, it can be very helpful to include cases that one knows well (the "Newark" cases) to assess whether one has sufficient bridges to anchor the response scale across cases.

5.4.3 Posterior predictive tests

Now imagine that you have a decent amount of bridging and that there are no islands or continents in your bridging network. Furthermore, you see no obvious face validity issues that raise suspicions that cases are insufficiently bridged. Are you done? The answer, unfortunately, is no. This problem is tricky because we simply do not know the true global scale of the data. But we can use a common Bayesian technique for model evaluation,

Figure 5.17: Effect of cross-state bridges on posterior predictive accuracy in the Yelp data. Panel A: distribution of posterior predictive probabilities for within-state versus cross-state reviews under a model fit *without* cross-state bridges, illustrating scale incompatibility. Panel B: the same distribution under the full model that includes cross-state bridges. When bridges are included, the within-state and cross-state distributions align closely.



posterior predictive checks, to probe for likely bridging failures.[33] Posterior predictive checks compare the data that we observe to what a fitted model predicts the data should look like. Here, we can take a fitted ordinal IRT model, use equation 5.6 to predict specific in-sample assessments, and then see how well we do. In particular, a well-fitting model should put relatively high probability on the respondent ratings that we actually observe.

While it generally makes sense to use posterior predictive checks to evaluate the performance of fitted models, certain sorts of observations are likely to prove most useful in the context of the bridge problem. For example, if our model does a worse job of predicting bridge observations than within-case observations, then one likely source of bad model fit is poor scale calibration across the respondents who assess different cases. On the other hand, if we see little difference in the predictive accuracy of the model across within-case and bridging observations, especially if that predictive accuracy is pretty high overall, then we can be reasonably confident that our model has estimated respondent thresholds in a way that consistently maps onto a single underlying latent scale. One needs bridge observations—preferably a non-negligible number of them—to apply this technique, of course. And it is not fool-proof. If we happen to obtain bridges from respondents who differ systematically from the respondents who do not provide bridges, we might see consistent predictive accuracy across bridge and within-case observations, but miss poor calibration in the respondents who do not provide bridges. This will be a particular concern if bridge observations tend to come from, or be directed to, particular observations. Therefore, posterior predictive tests can only serve as a companion to the other strategies that we describe in this section. If your bridge network is reasonably dense and well-distributed, the face validity of your cases is high, and you find that within-case and bridge observations have similar (high) posterior predictive accuracy, then you may be reasonably confident that your data are sufficiently bridged.

We first use the Yelp data to illustrate the potential value of this posterior predictive check approach. Figure 5.17 shows the distribution of individual posterior predictive probabilities, averaged within restaurants, comparing within-state to cross-state reviews.[34] Panel A shows what happens when bridges are excluded from the model: we fit a separate IRT model using only within-state reviews, then applied its estimated parameters to both within-state and cross-state ratings. The density of posterior predictive probabilities is shifted markedly downward for cross-state reviews (mean $p = 0.34$) compared to within-state reviews (mean $p = 0.41$), because the model cannot reconcile rating scales across states without bridge raters. Panel B shows the same comparison under the full model that includes cross-state bridges: the two distributions align closely (mean $p = 0.41$ within-state versus 0.41 cross-state), confirming that bridges are doing their intended job.

Panel A of Figure 5.18 disaggregates these results to the restaurant level, plotting mean within-state accuracy (x-axis) against mean cross-state accuracy (y-axis) for each restaurant with at least three cross-state reviews. The two dashed reference lines provide benchmarks for how far below the diagonal is cause for concern. The upper line marks a 20 percent relative drop in exact-match accuracy: a restaurant on this line has cross-state reviewers

[33] See a reference on Bayesian data analysis, such as Gelman et al. (2013), for a primer on posterior predictive checks.

[34] We define a Yelper's state as the state where most of their reviews appear.

whose ratings the model predicts 20 percent less often than it predicts within-state ratings, scaled to the dataset's average within-state accuracy of 0.43. Because exact-match is a strict criterion—being off by one category counts as a full miss—this line sits relatively close to the diagonal. The lower line applies the same 20 percent relative threshold to $1 - \text{RPS}$, a graded metric that gives partial credit for near-misses (see Appendix 5.B for a formal definition). Since the average within-state $1 - \text{RPS}$ is much higher (0.90), a 20 percent relative drop in RPS corresponds to a larger absolute gap between the two lines. A restaurant between the two lines has cross-state errors that are real by the strict exact-match standard but modest in magnitude—mostly off-by-one rather than large jumps across the scale. A restaurant below the lower line has cross-state errors that are large even when graded generously, a stronger signal of genuine scale mis-calibration. In the Yelp data, 32 percent of restaurants fall below the exact-match line and only 4 percent fall below the RPS line, suggesting that most cross-state errors in this dataset, while detectable, are relatively small in magnitude.

In Figure 5.18, node size represents how many bridge ratings—ingoing or outgoing—connect a restaurant to restaurants in other states. Shapes group together restaurants by how many incoming cross-state reviews they have. And shade indicates how discriminating the average cross-state reviewer for the restaurant is. We can use this information to identify cases that may need better bridging. First, small circles in Figure 5.18 are cases with few bridges overall. While the "right" bridge can anchor a case, it is fundamentally impossible to evaluate whether these cases are well-bridged, given the paucity of bridges. Second, cases well below the diagonal represent situations where the model does a better job of predicting in-state than out-of-state ratings. Such cases may be badly bridged. But there are different sorts of below-diagonal cases. A restaurant like The Butcher's Shop has only few, badly predicted, out-of-state ratings, from low-discrimination Yelpers. It is very hard to know, therefore, whether it is well below the diagonal because it is poorly bridged, or whether the bridges it does have are simply idiosyncratic. We probably want to collect more bridges for this case. On the other hand, restaurants like Hei La Moon, Pauli's, Taiwan Cafe, and Varasano's all have many bridges and a reasonable number of discriminating cross-state reviews. Here, we have pretty strong evidence that the model understands cross-state Yelpers less well than it understands in-state reviewers, likely providing evidence of a cross-state mis-calibration. Somewhere in the middle, we find the Daily Catch and Regina Pizzeria; these restaurants have numerous cross-state reviews and bridges, but the cross-state reviewers are relatively unreliable and may simply be hard to predict. Here, evidence for mis-calibration is more tentative.

Panel B of Figure 5.18 presents the equivalent diagnostic for V-Dem's *Political killings* variable. Here, experts are matched with individual countries and each V-Dem expert has a primary country. This contrasts with the Yelp data, where we define bridging by state, rather than by restaurant. Again, we see different sorts of suspect cases. For instance, China has few bridge coders—who provide few ratings of China—while Saudi Arabia is a well-bridged case that nonetheless exhibits substantially lower PPD accuracy for bridge coders than for experts who have Saudi Arabia as their main country. In both cases, the PPD check may indicate poor bridging. But they might also indicate other issues. China clearly needs more bridges, both to better anchor the case and to reliably assess the quality of that anchoring. Saudi Arabia has more bridge coders than main country experts. It is possible that these bridge coders exhibit highly variable understandings of the case, making them harder to

Figure 5.18: Posterior predictive diagnostics for cross-case bridging. Panel A: Yelp restaurants with at least three cross-state reviews (bridge model); mean posterior predictive probability for within-state reviews (x-axis) versus cross-state reviews (y-axis). Panel B: V-Dem *Political killings* (vignette model), equivalent plot at the country level. Panel C: Yelp restaurants, same as Panel A but using parameters estimated from the no-bridge model (within-state reviews only); cross-state reviews are evaluated out-of-sample. In all panels, points on the diagonal indicate equal predictive accuracy across local and bridge raters. Node size reflects the number of bridge raters, node shape the number of cross-case ratings, and node shading the average discrimination parameter (β) of bridge raters (darker = more discriminating). Dashed lines mark 20 percent relative accuracy drops by two criteria: exact-match probability (upper line) and 1 − RPS (lower line).

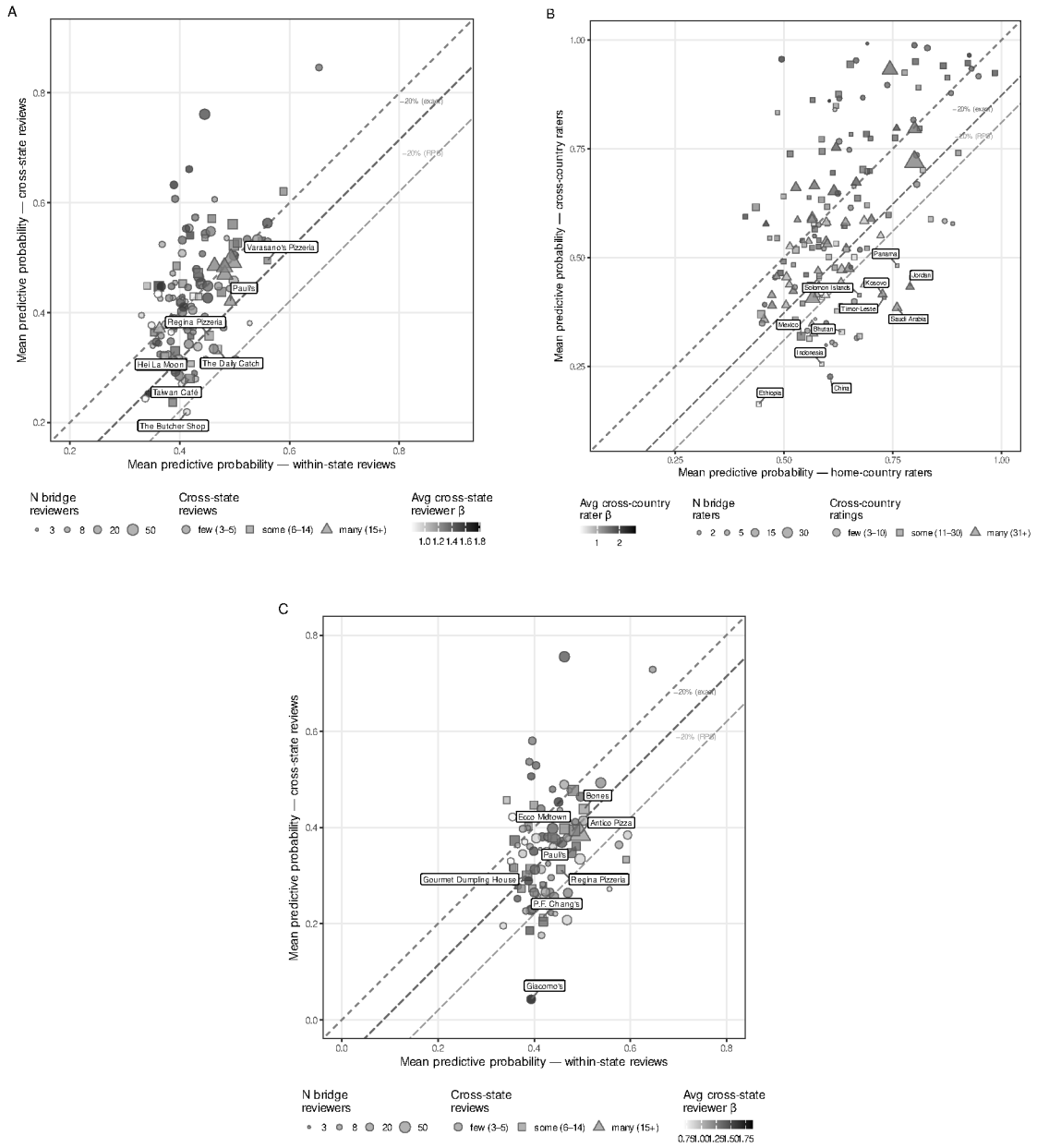


Table 5.3: Percentage of cases with cross-case predictive accuracy more than 10 percent or 20 percent below within-case accuracy (relative threshold, scaled to each dataset's mean within-case accuracy), under two accuracy metrics: exact-match probability and $1 - \text{RPS}$. "Yelp (no-bridge)" evaluates cross-state reviews out-of-sample using model parameters estimated on within-state reviews only.

	Yelp		Yelp (no-bridge)		V-Dem	
	$> 10\%$	$> 20\%$	$> 10\%$	$> 20\%$	$> 10\%$	$> 20\%$
Exact-match accuracy	31.6%	18.0%	65.3%	48.4%	46.7%	30.5%
$1 - \text{RPS}$	3.8%	0.0%	15.8%	2.1%	4.2%	0.6%
Mean within-case (exact)	0.43		0.43		0.64	
Mean within-case ($1 - \text{RPS}$)	0.90		0.90		0.95	
N (cases)	133		95		167	

predict. In any case, this sort of well-identified mis-calibration is clearly an issue worthy of further investigation.

For reference, Panel C of Figure 5.18 shows what this plot looks like for a fully un-bridged Yelp dataset. Now, almost no cases predict cross-state ratings as well as they do in-state ratings, and we see numerous cases falling below our rule-of-thumb warning thresholds. Along with Figure 5.17, this visualization helps illustrate the overall value of this type of PPD check for finding bridging and anchoring problems.

Stepping back to view the forest, instead of the trees, Table 5.3 shows that the severity of cross-case degradation depends on the dataset. In the Yelp data, where all raters are bridge raters and baseline accuracy is modest (0.43), roughly 32 percent of restaurants fall below the 10 percent exact-match accuracy threshold and 18 percent below 20 percent. In V-Dem, where cross-country raters are a minority and baseline accuracy is higher (0.64), the proportions are larger: 47 percent and 31 percent, respectively. That said, neither dataset shows a substantial number of cases with poor performance on the 1-RPS diagnostic. When we assess mis-classification using the full posterior probability distribution, and weight mis-classification by distance from the true value, then we find only a few cases that show a clear drop-off in predictive performance. That said, we see substantial mis-classification, even using the 1-RPS diagnostic, when making out-of-sample predictions with the un-bridged Yelp dataset, again demonstrating the potential for these diagnostics to uncover poor bridging.[35]

As you can see, using PPD checks to diagnose bridging issues is more of an art than a science. But this example provides some guidance in what to look for. Well-bridged cases for which the model is bad at predicting discriminating bridge ratings are a red flag. Bad prediction for lightly, or unreliably, bridged cases provides a less clear signal, but one should collect more—and better—bridges for such cases, to be safe. Plots like those in Figures 5.17 and 5.18 provide a bird's eye view of the overall potential for bridging failure, while helping

[35]Of course the cross-state predictions are out-of-sample in this exercise, which may explain some of the poor predictive accuracy. Nonetheless, in a fully bridged scenario, we would be able to accurately predict out-of-state raters' case ratings, out of sample, based on the in-state scores and the out-of-state respondents' estimated threshold parameters.

one to identify particularly suspect cases.

5.5 Conclusion

In this chapter, we engaged a core problem in making surveys comparable across cases: bridging. Bridging issues are endemic in comparative studies in the social sciences, and even on ratings sites. They are also difficult to solve. It can be quite difficult to ask respondents to answer questions about multiple cases—indeed, respondents and cases are often inexorably linked. As we show in this chapter, ignoring the bridging problem can lead us to draw incorrect conclusions. Therefore, we argue that a researcher should carefully examine their data for bridging issues by asking if respondents are isolated by case (or subset of cases), by examining face validity with bridging issues in mind, and by using posterior predictive tests. We show that one can be strategic when addressing bridging issues by working to obtain bridges that are most informative, and we provide a few rules of thumb. First, learn as much about respondent thresholds as possible. This means one should prioritize obtaining bridge ratings from respondents who evaluate uninformative (e.g., "static") cases. Second, make sure to bridge to and from as many cases as you can. Finally, tools like the MELPD algorithm can help a researcher strategically select bridges when intuition fails.

5.A Variational Approximation for O-IRT

Variational approximation algorithms allow one to approximately estimate an intractable density function, $p(\boldsymbol{\theta}|\mathbf{y})$, by iteratively approximating the target density in terms of a more tractable density, $q(\boldsymbol{\theta})$, in a way that minimizes the Kullback-Leiber divergence between q and $p(\cdot|\mathbf{y})$. In a common approach, known as a mean field approximation, one selects an approximating density q from a parametric family such that, for some parameter partition $\{\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_M\}$, $q(\boldsymbol{\theta})$ factorizes into $\prod_{i=1}^{M} q_i(\boldsymbol{\theta}_i)$ (Ormerod & Wand 2010). Interestingly, Ormerod & Wand (2010) show that mean field approximation and Gibbs sampling are closely related. In particular, they show that the optimal sub-densities $q_1, \ldots, q_M$ for minimizing the Kullback-Leiber divergence between q and p are

$$q_i^*(\boldsymbol{\theta}_i) \propto \exp\left[E_{\neg\boldsymbol{\theta}_i} \log p(\boldsymbol{\theta}_i|\boldsymbol{\theta}_{\neg i}, \mathbf{y})\right]. \tag{5.7}$$

In other words, these optimal approximating distributions are functions of the full conditional distributions of the parameters of p. Thus, it is generally straightforward to adapt any Gibbs sampling algorithm, which simulates from p by iteratively drawing from the full conditional distributions of p, into a mean field approximation algorithm. This is helpful in this context because researchers have extensively developed Gibbs sampling algorithms for IRT models (see e.g. Johnson & Albert 1999).

5.A.1 A mean field approximation algorithm for the O-IRT model

We develop a mean approximation algorithm for the O-IRT model, subject to the following priors:

$$z_{ct} \sim \mathcal{N}(0,1), \beta_r \sim \mathcal{N}(\mu_\beta, \sigma_\beta^2), \text{and } \gamma_{r,k} \sim U(-2,2) \text{ s.t } \gamma_{r,1} \leq \cdots \leq \gamma_{r,K-1}.$$

In particular, we construct an approximating distribution, q, such that

$$q(\boldsymbol{\theta}) = q_{\tilde{y}_{ctr}}(\tilde{y}_{ctr}) q_{z_{ct}}(z_{ct}) q_{\beta_r}(\beta_r) q_{\gamma_{r,k}}(\gamma_{r,k})$$

where $\boldsymbol{\theta}$ is the full vector of model parameters, and the optimal approximating partitioned distributions are

$$
\begin{aligned}
q^*_{\tilde{y}_{ctr}}(\tilde{y}_{ctr}) &\propto \exp\left[E_{\neg\tilde{y}_{ctr}} \log\ p(\tilde{y}_{ctr}|\boldsymbol{\theta}_{\neg y_{ctr}}, y_{ctr})\right] \\
&\sim \mathcal{TN}\left(E(z_{ct})E(\beta_r), 1, E(\gamma_{r,y_{ctr}-1}), E(\gamma_{r,y_{ctr}})\right), \\
q^*_{z_{ct}}(z_{ct}) &\propto \exp\left[E_{\neg z_{ct}} \log\ p(z_{ct}|\boldsymbol{\theta}_{\neg z_{ct}}, \mathbf{y}_{ct})\right] \\
&\sim \mathcal{N}\left(\frac{\sum_{r\in R_{ct}} E(\beta_r)E(\tilde{y}_{ctr})}{1+\sum_{r\in R_{ct}} E({\beta_r}^2)}, \frac{1}{1+\sum_{r\in R_{ct}} E({\beta_r}^2)}\right), \\
q^*_{\beta_r}(\beta_r) &\propto \exp\left[E_{\neg\beta_r} \log\ p(\beta_r|\boldsymbol{\theta}_{\neg\beta_r}, \mathbf{y}_r)\right] \\
&\sim \mathcal{N}\left(\frac{\sum_{c,t\in J_r} E(z_{ct})E(\tilde{y}_{ctr}) + \frac{\mu_\beta}{\sigma^2_\beta}}{\frac{1}{\sigma^2_\beta}+\sum_{c,t\in J_r} E({z_{ct}}^2)}, \frac{1}{\frac{1}{\sigma^2_\beta}+\sum_{c,t\in J_r} E({z_{ct}}^2)}\right), \\
q^*_{\gamma_{r,k}}(\gamma_{r,k}) &\propto \exp\left[E_{\neg\gamma_{r,k}} \log\ p(\gamma_{r,k}|\boldsymbol{\theta}_{\neg\gamma_{r,k}}, \mathbf{y}_r)\right] \\
&\sim U\left(\max\left[-2, \max_{y_{ctr}=k} E(\tilde{y}_{ctr})\right], \min\left[\min_{y_{ctr}=k+1} E(\tilde{y}_{ctr}), 2\right]\right).
\end{aligned}
\tag{5.8}
$$

The expectations in the approximating distributions are

$$
\begin{aligned}
E(\tilde{y}_{ctr}) &= \mu_{ctr} + \frac{\phi(E(\gamma_{r,y_{ctr}-1}) - \mu_{ctr}) - \phi(E(\gamma_{r,y_{ctr}}) - \mu_{ctr})}{\Phi(E(\gamma_{r,y_{ctr}}) - \mu_{ctr}) - \Phi(E(\gamma_{r,y_{ctr}-1}) - \mu_{ctr})} \\
E(z_{ct}) &= \frac{\sum_{r\in R_{ct}} E(\beta_r)E(\tilde{y}_{ctr})}{1+\sum_{r\in R_{ct}} E({\beta_r}^2)} \\
E(\beta_r) &= \frac{\sum_{c,t\in J_r} E(z_{ct})E(\tilde{y}_{ctr}) + \frac{\mu_\beta}{\sigma^2_\beta}}{\frac{1}{\sigma^2_\beta}+\sum_{c,t\in J_r} E({z_{ct}}^2)} \\
E(\gamma_{r,k}) &= \frac{1}{2}\left(\max\left[-2, \max_{y_{ctr}=k} E(\tilde{y}_{ctr})\right] + \min\left[\min_{y_{ctr}=k+1} E(\tilde{y}_{ctr}), 2\right]\right)
\end{aligned}
$$

where $\mu_{ctr} = E(z_{ct})E(\beta_r)$. Note that the notation $E(x)$ is shorthand for the expected value of x with respect to the approximating distribution $q_x(x)$. One confusing aspect of these equations is that they appear to be infinitely recursive. So, for example, $E(\tilde{y}_{ctr})$ is a function of $E(z_{ct})$ which is a function of $E(y_{ctr})$. But, at any point in the algorithm, the current estimate of each $q_x(x)$ distribution is fixed, and we have an approximation to $E(x)$ for each x that can be plugged directly into whatever distribution function we are updating our estimate at the given step.

The algorithm works by iteratively approximating the values $E(y_{ctr}), E(z_{ct}), E(\beta_r)$, and $E(\gamma_{r,k})$ across c, t, r, and k. Upon convergence, which we evaluate by monitoring change in the lower bound on the marginal likelihood

$$
\underline{p}(\mathbf{y}, q) \equiv \exp \int q(\boldsymbol{\theta}) \log\left[\frac{p(\mathbf{y}, \boldsymbol{\theta})}{q(\boldsymbol{\theta})}\right] d\boldsymbol{\theta},
$$

we can calculate approximate point estimates and credible intervals for parameters of interest by plugging our final estimates of these expectations into the approximating conditionals described in equation 5.8.

5.B The Ranked Probability Score

The Ranked Probability Score (Epstein 1969) is a proper scoring rule for ordinal or probabilistic forecasts. For a rating with K ordered categories, let $p_k = \Pr(y = k \mid \boldsymbol{\theta})$ be the model's predicted probability for category k, and define the predicted cumulative distribution function as $F_k = \sum_{j=1}^{k} p_j$. Let $O_k = \mathbf{1}(y_{\text{obs}} \leq k)$ be the corresponding step function for the observed outcome. The RPS is then

$$\text{RPS} = \frac{1}{K-1} \sum_{k=1}^{K-1} \left(F_k - O_k\right)^2 . \tag{5.9}$$

RPS equals zero when all probability mass is placed on the observed category and reaches its maximum of one when all mass is placed on the category furthest from the observed outcome. Unlike exact-match accuracy, RPS penalizes predictions in proportion to how far they are from the observed value: being off by one category incurs a smaller penalty than being off by three.

We report $1 - \text{RPS}$ so that higher values indicate better predictive accuracy, on the same intuitive scale as exact-match probability. For a five-category scale ($K = 5$), a perfect prediction yields $1 - \text{RPS} = 1$; a prediction that places all mass one category away from the observed value yields approximately $1 - \text{RPS} = 0.75$.