

Evaluating and Improving Item Response Theory Models for Cross-National Expert Surveys*

Daniel Pemstein[†], Eitan Tzelgov[‡], Yi-Ting Wang[§]

Abstract

The data produced by the Varieties of Democracy (V-Dem) project contains ordinal ratings of a multitude of country-level indicators across space and time, with multiple experts providing judgments for each country-year observation. We use an ordinal item response theory (O-IRT) model to aggregate multiple experts' ratings. The V-Dem data provide a challenging domain for such models because they exhibit little cross-national *bridging*. That is, few coders provide ratings for multiple countries, making it difficult to calibrate the scales of estimates cross-nationally. In this paper, we provide a systematic analysis of the issue of *bridging*. We first use simulations to explore how much bridging one needs to achieve scale identification when coders' thresholds vary across countries and when the latent traits of some countries lack variation. We then examine how posterior predictive checks can be used to check cases of extent of scale non-comparability. Finally, we develop and evaluate search algorithms designed to select bridges that are most likely allow one to correct scale incompatibility problems.

*The authors would like to thank Andrew Martin, Kevin Quinn, and the other members of the V-Dem team for helpful suggestions. This research was supported, in part, by National Science Foundation Grant SES-1423944 and Riksbankens Jubileumsfond Award M13-0559:1.

[†]Assistant Professor, North Dakota State University, daniel.pemstein@ndsu.edu.

[‡]Postdoctoral Research Fellow, University of Gothenburg, eitan.tzelgov@gu.se.

[§]Assistant Professor, National Cheng Kung University, yitingw@mail.ncku.edu.tw

Latent concepts are ubiquitous in political science. The study of political systems, policies, and belief systems requires careful theorizing and exploration of concepts that are not directly measurable. In particular, constructs such as level of democracy, extent of human rights, and ideology are all latent qualities that we cannot observe directly. Nonetheless, social scientists have theoretical reasons to work with measures of such latent constructs and core questions in our disciplines are virtually impossible to tackle if we restrict the objects of our analyses to concrete observables. Therefore, we often rely on the subjective judgment of experts to provide measures of latent traits.

Of course, there are many ways to leverage expert knowledge for concept measurement. One approach places substantial faith in a small number of experts. For instance, one might use one specialist, or a small panel of judges, who jointly code the set of observations, say politics (Marshall & Gurr 2011). This strategy assumes that a small group of experts have access to the knowledge and information necessary to adequately measure latent traits across the population of interest. Not surprisingly, this assumption may be difficult to meet in cross-national studies where detailed local knowledge is necessary to adequately assess concepts. Thus, in cross-national research, it is common to distribute the load over a large number of experts, and ask each specialist to code a single unit, say a party's election manifesto (Klingemann, Volkens, Bara, Budge & McDonald 2006). Ideally, this strategy draws from a cross-national pool of experts that can leverage local knowledge to produce detailed and valid measures.

This measurement strategy relaxes the assumption that a small group of experts can evaluate many units, but has the potential to generate differential assessments across units because it implicitly assumes that experts' understandings of concepts are identical. In other words, these approaches simply assume that each coder is able to produce a valid point-estimate of the latent concept at hand.¹ Multiple raters may have quite different understandings of latent concepts and this variation in conceptualization may be especially

¹Klingemann et al. (2006) are a somewhat inappropriate example here, because they attempt to train their raters to maximize inter-coder reliability.

exacerbated when experts hail from varying cultural and educational traditions. Indeed, in the context of party manifestos, König, Marbach & Osnabrugge (2013) show that expert ratings exhibit substantial cross-national bias. More generally, people with varying backgrounds are likely to apply different standards when rating concepts (King & Wand 2006). For example, consider the task of rating the extent of media freedom in a country, using a Likert scale running from zero (no freedom) to five (complete freedom). If experts in country A—perhaps because of a long history of freedom of expression—hold media freedom to higher standards than those in country B, then country A expert thresholds will be higher than those of their counterparts. Even within countries, experts standards may differ. And, varying standards aside, research shows that people’s responses to Likert scales vary systematically across cultures (Lee, Jones, Mineyama & Zhang 2002). Thus, measurement efforts that rely on cross-national teams of experts need to take the possibility of varying standards seriously.

Both of the aforementioned strategies place substantial faith in each individual expert, and—because they generate only one estimate per observation—provide little leverage over potential measurement uncertainty. In essence, therefore, they assume that the expert ratings that they rely upon are error-free, or at least highly reliable.² Expert surveys, where more than one expert provides a rating of each observation, allow researchers to relax this reliability assumption and, at least potentially, produce estimates of measurement error (Bakker, de Vries, Edwards, Hooghe, Jolly, Marks, Polk, Rovny, Steenbergen & Vachudova forthcoming, Fish & Kroenig 2009, Kitschelt 2013). With these approaches, however, aggregating coders’ opinion becomes an issue. In most cases these studies focus on measures of central tendency, and ignore precision, or, at best, provide simple summaries of variation in expert codes. Moreover, this approach does little to alleviate concerns about cross-expert equivalency; there is no way of knowing whether experts have the same scale in mind when they provide their codings. For example, we do not know if a Hungarian specialist defines

²Although many of these projects do attempt to verify the inter-coder reliability of their experts or coding teams.

extreme-right political parties in the same way that a Swedish expert might. While expert surveys typically use carefully worded questions that are designed to minimize conceptual slipping, there remains a substantial risk that experts in different contexts might interpret coding rules in systematically different ways.

A more sophisticated approach to aggregating expert ratings explicitly models the measurement process, typically using item response theory (IRT) methods developed by scholars of educational testing (see e.g. Clinton & Lapinski 2006, Clinton & Lewis 2007, Treier & Jackman 2008, Pemstein, Meserve & Melton 2010, Linzer & Staton 2012, König, Marbach & Osnabrügge 2013, Schnakenberg & Fariss 2014, Fariss 2014). These models relax two key assumptions implicitly made by traditional approaches to aggregating ratings in expert surveys. First, they allow for variation in rater precision. Second, they allow raters to adopt differing standards of conceptualization. In particular, when raters provide ordinal (Likert scale) ratings of a particular concept, these methods allow for the possibility that different experts have varying thresholds for placing observations into particular categories within the scale. Finally, because IRT methods are grounded in an explicit model of the rating process, they both aggregate information provided by multiple experts and produce reasoned estimates of uncertainty around aggregate point estimates.

However, these tools are not immune to the problems that plague other strategies for measuring latent social science concepts cross-nationally. In particular, while these methods can, in principle, deal with variation in how experts perceive latent traits, common patterns of data collection in comparative politics—in particular, using disjoint sets of coders across countries—make it difficult for researchers to use IRT models to produce cross-nationally comparable measures of latent concepts. Thus, a persistent problem in the use of these methods has been the issue of calibration/comparability, or lack of *bridging*. Since in many cases the same latent concept is being measured for different units and/or in different times, it is vital to identify *bridges* who cross disjoint sets of coders and provide information necessary to calibrate coder thresholds across units. In our context, these are experts who are capable

of rating more than one country. This problem is analogous to that faced by researchers who use IRT methods to estimate ideology from roll call votes and who wish to create *common space* measures across institutions (McCarty & Poole 1995, Bailey 2007, Shor & McCarty 2011) or within institutions across time (Poole & Rosenthal 1997, Groseclose, Levitt & Snyder 1999, Martin & Quinn 2002). Yet, we have little understanding of exactly how much bridging is necessary to achieve scale equivalency across units, and recent work has called into question the extent to which existing common spaces actually achieve scale identification (Ho & Quinn 2011).

In this paper we provide a systematic analysis of the issue of *bridging*, motivated by the *Varieties of Democracy* (V-dem) project. This project is, at its core, an expert-survey, in which country experts provide ordinal ratings about dozens of democratic indicators (e.g. barriers to parties, freedom of academic speech, civil society participation), for the 1900-2012 period. The unit of analysis for each indicator is country-year, for which at least five coders provide ratings. A major selling point of the project is its reliance on a mix of local experts and foreign specialists; generally about half of the raters reside in the country that they rate. But this focus on specialized knowledge comes at a cost; initial survey waves almost exclusively asked experts to provide ratings for only a single country and the length of the survey made both the creation and use of anchoring vignettes (King & Wand 2006) impractical. Subsequent waves have focused on ‘bridge-coding,’ but the search for bridge coders faces a number of practical limitations. First, the specificity of the survey³ means that few potential raters have the ability to provide cross-national coding, especially for the whole time period under consideration. Resources are also an important constraining factor because recruiting bridge-coders and administering surveys is expensive. Thus the V-Dem project faces a problem that is—or should be—generally applicable to a large class of cross-national expert surveys: it needs to recruit bridge coders as efficiently as possible and to evaluate when bridge coding is sufficient to provide cross-nationally comparable estimates of

³Technically, surveys. V-Dem experts are partitioned into over ten specializations and each set of experts completes a survey specific to her area of expertise.

latent traits.

Thus, we seek to answer a set of key questions. First, if we adopt weak prior assumptions about the structure of rater thresholds,⁴ how much bridging is necessary to achieve cross-national comparability in a survey structured like V-Dem? Perhaps more importantly, how do we know when we have enough bridges? Or, failing that, can we at least develop tools to systematically identify cases for which scale comparability is likely to be a problem? Finally, how can we most efficiently select bridges in order to obtain scale comparability across countries? Or, in other words, what set of bridges are likely to provide such comparability at the lowest cost to the project?

In the next section we describe the IRT model at the heart of this project. Next, we illustrate the bridging problem in V-Dem and then use simulations to explore how much bridging one needs to achieve scale identification when faced with two types of problems that seem common in V-Dem, and probably plague many cross-national expert surveys of social science concepts. We then examine how useful posterior predictive checks are for identifying instances of scale non-comparability. Finally, we develop and evaluate a pair of search algorithms designed to select bridges that are most likely to provide information, and thus allow one to correct for, scale incompatibility problems.

1 An O-IRT Model for V-Dem Data

We need to introduce some notation to keep track of things because the complexity of the V-Dem dataset, which contains ratings of a vast number of indicators across space and time. Moreover, each indicator is rated by more than one judge. This means that there are

- $i \in I$ indicator variables,

⁴Weak priors allow raters to have widely varying thresholds. Very strong priors would assume, as most projects involving cross-national expert surveys essentially do, that thresholds are identical across raters. Middle-ground assumptions might maintain that same-country experts have identical, or very similar, thresholds, or that raters with similar backgrounds—education, profession, and so forth—have similar thresholds to one another.

- $r \in R$ raters,
- $c \in C$ countries,
- and $t \in T = \{1, \dots, \bar{t}\}$ time periods.

I is the set of indicator variables while i represents one element from that set, and so on. Each of the $|R|$ raters provides ratings of one or more of each of the $|I|$ indicators in some subset of the available $n = |C| \times |T|$ country-years covered by the dataset. Each country enters the dataset at time \underline{t}_c and exits at time $\bar{t}_c + 1$. We'll call rater r 's set of observed ratings/judgments J_r . Each element of each of these judgment sets is an i, c, t triple. Similarly, the set of raters that rated country-year c, t is R_{ct} . In this paper we focus on models for a single indicator, and therefore drop the i indices from our notation. We also restrict our attention to the case where indicator i is ordinal.

For a given indicator we observe a sparse⁵ $|C| \times |T| \times |R|$ array, \mathbf{y} , of ordinal ratings. While our observations are ordinal, we assume that raters make judgments about latent continuous concepts. In particular, we assume that each rater first perceives latent values with error, such that

$$\tilde{y}_{ctr} = z_{ct} + e_{ctr} \tag{1}$$

where z_{ct} is the “true” latent value of the given indicator in country c at time t , \tilde{y}_{ctr} is rater r 's perception of z_{ct} , and e_{ctr} is the error in rater r 's perception for the country-year observation. For the cases that she judges (J_r), rater r places a country-year in category k if $\tau_{r,k-1} < \tilde{y}_{ctr} \leq \tau_{r,k}$, where each τ represents a rater cutoff point on the underlying latent scale. The vector $\boldsymbol{\tau}_r = (\tau_{r,1}, \dots, \tau_{r,K-1})$ is the vector of unobserved ranking cutoffs for judge r on the latent scale. We fix each $\tau_{r,0} = -\infty$ and $\tau_{r,K} = \infty$, where K is the number of ordinal categories raters use to judge the indicator. For the moment, say the cumulative

⁵The majority of raters provide ratings for only one country.

distribution function for the rating errors is

$$e_{ctr} \sim F(e_{ctr}/\sigma_r). \quad (2)$$

Combining this assumption, and the assumptions about rater cutoffs described in the above paragraph, with equation 1 implies the following data generating process (DGP):

$$\begin{aligned} \Pr(y_{ctr} = k) &= \Pr(\tilde{y}_{ctr} > \tau_{r,k-1} \wedge \tilde{y}_{ctr} \leq \tau_{r,k}) \\ &= \Pr(e_{ctr} > \tau_{r,k-1} - z_{ct} \wedge e_{ctr} \leq \tau_{r,k} - z_{ct}) \\ &= F\left(\frac{\tau_{r,k} - z_{ct}}{\sigma_r}\right) - F\left(\frac{\tau_{r,k-1} - z_{ct}}{\sigma_r}\right) \\ &= F(\gamma_{r,k} - z_{ct}\beta_r) - F(\gamma_{r,k-1} - z_{ct}\beta_r). \end{aligned} \quad (3)$$

The last two lines of equation 3 reflect two common parameterizations of this model. The first parameterization is typically called multi-rater ordinal probit (MROP) (Johnson & Albert 1999, Pemstein, Meserve & Melton 2010),⁶ while the latter is an ordinal item response theory (O-IRT) setup (Clinton & Lewis 2007, Treier & Jackman 2008). Note, in particular, that $\beta_r = \frac{1}{\sigma_r}$ and $\gamma_{r,k} = \frac{\tau_{r,k}}{\sigma_r}$.⁷ The parameter σ_r is a measure of rater r 's reliability when judging the indicator; specifically it represents the size of r 's typical errors. Raters with small σ_r parameters are better, on average, at judging indicator i than are raters with large σ_r parameters. In the IRT literature, β_r is known as the discrimination parameter,⁸ while each γ is a difficulty parameter. The discrimination parameter is a measure of precision; a rater characterized by an item discrimination parameter close to zero will be largely unresponsive to true indicator values when making judgments while a rater with a discrimination parameter far from zero will be very “discriminating.” The γ and τ parameters are thresholds that control how raters map their perceptions on the latent interval scale into ordinal

⁶If we assume $F(\cdot)$ is standard normal.

⁷This equivalency breaks down if we allow for β_r parameters less than one. Thus, the O-IRT model is potentially more general than MROP.

⁸Equation 3 makes clear that β_r measures rater precision.

classifications. The term “difficulty parameter” stems from applications in educational testing where the latent variable is ability and observed ratings are binary (in)correct answers to test questions. The scale identification problem faced by cross-national surveys such as V-Dem is driven by the fact that these γ and τ parameters may—and perhaps are even likely to—vary across raters hailing from different cultural and educational backgrounds.

Completing the model specification requires adopting prior distributions for the model parameters. We will focus on the O-IRT model here. Generally, one sets $z_{ct} \sim \mathcal{N}(0, 1)$ a priori. This assumption (arbitrarily) sets the scale of the estimated latent traits. Second, one assumes $\beta_r \sim \mathcal{TN}(\mu_{\beta_r}, \sigma_{\beta_r}, 0, \infty)$.⁹ Finally, while different authors adopt varying priors for $\gamma_{r,k}$, it is possible to adopt completely uninformative uniform priors for these parameters, subject to the constraint that they remain ordered within raters. When one lacks any bridging between sets of raters, such a prior will fail to identify the model. Or, in other words, under such a prior, the model lacks sufficient information to set the relative placement of disjoint (with respect to cases rated) sets of rater thresholds, and, in turn, latent trait estimates. Thus, we adopt vaguely informative priors for the $\gamma_{r,k}$ parameters throughout.

At its heart, this model does three things. First, it takes ordinal observations and maps raters’ thresholds onto a single interval-valued latent variable. In other words, it provides a reasoned way to deal with a relatively large class of differences in how individual respondents interpret Likert scales. Second, it allows raters to vary in how reliably they make judgments, but largely assumes away the potential for systematic rater biases that are not covered by varying thresholds.¹⁰ This latter point is easiest to see in the MROP version of the model. Specifically, in a standard MROP, one assumes $F(\cdot)$ is standard normal, such that $e_{ctr} = \mathcal{N}(0, \sigma_r^2)$. So, in other words, raters get things right on average, but they make

⁹1 is a common choice for μ_{β_r} . One typically chooses a large value for σ_{β_r} , thereby adopting a vague prior. The assumption of truncation at zero equates to assuming that raters agree on which direction is up. While reasonable in our context, this assumption is inappropriate when one uses IRT models to estimate ideology from voting behavior.

¹⁰For instance, the model cannot account for a rater that applies one set of thresholds to one country and a different set to another. Nor does this model capture the possibility that rater precisions might vary over space and time, although the model might be expanded to handle such issues.

stochastic mistakes where the typical magnitude of mistakes that rater r makes on indicator i is σ_r^2 . So, if $\sigma_r^2 < \sigma_{r'}^2$, then rater r provides more reliable judgments about \mathbf{z} than r' because she makes smaller mistakes on average. Finally, taking differences in rater thresholds and precisions into account, the model produces interval-valued estimates of latent traits—each z_{ct} —accompanied by estimates of measurement error that reflect both the level disagreement between coders on the case in question, and the estimated precision of the coders who rated the case.

2 The Bridge Problem

In practice, we face two related problems when attempting to use O-IRT models to generate cross-nationally comparable estimates of latent traits using the V-Dem data. The general problem, of course, is that, if we choose not to simply assume that raters share similar thresholds,¹¹ we cannot establish the cross-national comparability of our estimates. And, because the scale of the model is established by the vague prior on the z_{ct} parameters, we know that, in the absence of any bridging, the scale location of scores within countries will be set relative to the empirical pattern of codes provided by raters. If country traits are highly variable and generated randomly from the same distribution, then scales would tend to converge even without bridging, but real traits do not look remotely like a random, highly variable, process. Furthermore, whether or not bridging observations establish scale convergence relies not just on whether or not we have many bridges, but on how bridging observations correspond to empirical patterns in the data. In other words, some bridges are more informative than others and certain countries are much harder to align with a common scale than others. Indeed, our second primary concern is that empirical patterns in regime traits often exacerbate the fundamental scale equivalence problem.

While it is, in general, difficult, to determine whether two cases are sufficiently bridged to

¹¹And fitting O-IRT models to single-country series provides substantial evidence that rater thresholds vary, even within countries.

obtain scale equivalence, sometimes the lack of sufficient bridging is painfully obvious. When we naively apply O-IRT models to weakly bridged V-dem data, we consistently produce estimates of latent values for Western-European countries that are lower than for countries in which experts codings’ on the original Likert scales point to serious democratic deficiencies. This finding is largely a result of two trends in our data. First, coders of Western, established democracies exhibit low variation in their coding. An example of this issue is presented in Figure 1. The top two panels present the yearly average ratings for Denmark on the original ordinal scale for V-Dem’s barriers to parties indicator. Country experts clearly agree that there are no barriers to parties in Denmark (higher scores reflect lower barriers), except during a brief period in the 1940s. Contrast this result with the highly variable yearly average of this variable for Mali, shown in the top right panel of the figure. The two bottom panels present the result of fitting an O-IRT model, like that described in section 1, to around 150 countries worth of ratings containing few cross-national bridges. While the measurement model results for Mali reflect its transition to a “no barriers” mode (as the z_{ct} estimates cross the upper threshold),¹² the measurement model estimates for Denmark never cross the upper threshold. Thus, even though the country experts agree that the situation with regards to parties’ barriers is flawless in Denmark, the lack of variation in ratings is problematic. Clearly, we lack sufficient bridges to place Denmark and Mali on the same scale. Moreover, a reasonable interpretation of the Danish data is that, because every rater agrees that Denmark rates very highly on this trait in every time period, we have should place Denmark near the top of the common scale with high confidence. But, once we relax the assumption that thresholds have consistent meanings across raters, country time-series like Denmark provide virtually no information about scores, because we have no way to nail down what Danish thresholds mean relative to one another, or to coders in other countries.¹³ Furthermore, while cases that exhibit sufficient variation to produce estimates

¹²Thresholds in these graphs reflect population means for the $\gamma_{r,k}$ parameters.

¹³This problem is akin to the issue pointed out by Spirling and McLean (2007), who argue that the optimal classification algorithm places extreme members of the UK parliament at the center of the ideological spectrum because of perfect ideological voting of these MPs.

that span their local scales may look comparable on visual inspection, without sufficient bridging it is impossible to establish the comparability of country time-series, even when they provide substantial information about coder thresholds within countries.

[Figure 1 about here]

3 A Simulation-Based Examination of Bridging

3.1 Thresholds Drawn from the Same Distribution

To evaluate the effects of bridging on model fit, cross-national comparability and identification, we preform a number of Monte Carlo experiments. Our first set of experiments examines the “Swiss,” or constant-country problem. These experiments, which examine scale equivalence between two simulated “countries,” proceed as follows:

1. We generate “true” values for each z_{ct} , which represent real country-year latent ability scores. In order to proxy the countries we observe in the V-dem dataset, we focus on two country ‘types.’ The first is what we call a ‘constant’ country (used as proxies for Western-European countries). To simulate these countries, we draw latent ability scores $z_{ct}^{constant} \sim \mathcal{N}(1.8, .05)$. In addition, we generate latent ability scores for countries we dub ‘random.’ These are countries in which coders’ ratings exhibit a high degree of variation, and thus provide us with a substantial information regarding coders’ thresholds. These scores are drawn from a $z_{ct}^{random} \sim \mathcal{N}(0, 1)$ distribution.¹⁴
2. We also generate $|R| \times K$ coders’ thresholds, $\gamma_{r,k}$. The first threshold, $\gamma_{r,1}$, is drawn from a $\mathcal{N}(-1.5, 0.2)$ distribution. Additional three, $\gamma_{r,k=(2,\dots,4)}$ are drawn sequentially from normal distributions with means $(-5., .5, 1)$ and standard deviation 0.2. For each threshold, we truncate the prior from below by the mean of the $k - 1$ threshold.

¹⁴Throughout the paper we use the term *constant country* to denote the country whose latent ability scores were drawn from a $\mathcal{N}(1.8, .05)$ distribution, and ‘random country’ to denote a $\mathcal{N}(0, 1)$ country. Random countries are, of course, an ideal type never actually realized in the V-Dem data.

Note that, while we allow rater thresholds to vary, the fact that we draw them from identical distributions will tend to minimize the role that such variation plays in scale equivalence. Thus, we focus attention on the issue that constant countries pose in these experiments.

3. To complete the latent data, we generate coders’ discrimination parameters such that $\beta_r \sim \mathcal{TN}(1, 3, 0, inf)$ distribution.
4. Given the above simulated parameters, we use equation 3 to simulate observed ratings data.

For our purposes, bridge-coding is defined as a situation where a coder who ‘originally’ codes one country (for example, an expert of British politics, rating only the UK), also provides ratings for an additional country. As for our stylized, simulated datasets, we generally begin with ‘baseline’ data, in which we simulate data for two countries, over a period of 100 years. This baseline dataset includes five coders who provide ratings *only* for one country, and five who provide ratings for the other. This is a ‘no-bridging’ scenario. In order to simulate bridge coding, we choose a number of coders from only *one* country, and simulate their ratings for the other. We do this gradually, starting with only one bridge coder, and up until we have ‘full-bridging.’ This is a situation where all five coders from a given country also provide ratings for the other.

We begin our evaluation of the effects of bridging on the mean square error (MSE) between the model’s estimates and the real latent scores. Bridge-coding is simulated for the entire period (100 years). The results of these simulations are shown in Figures 2 and 3. The figure is divided into sub-regions, capturing the different levels of simulated bridge coding (e.g., the left sub-region is a ‘no-bridging’ scenario, and in the fourth, there are three coders from the first country who also rate the second). We replicated each experiment across ten simulated datasets. The figures demonstrate that full time-series bridging, done either from the constant to the random country or vice versa, results in a rapid improvement in terms of

MSE. Most strikingly, once we simulate one bridge coder, the MSE values for both country types decrease from approximately 1 to 0.3 (in the constant country), and approximately 0.85 to 0.2 (random). Once we add more bridge coders MSE values are further reduced, and reach a plateau of approximately 0.15-2 for constant countries and 0.1-0.15 for the random country with 3-5 bridge coders.

[Figures 2, 3 about here]

To assess the impact of the quantity of bridge coding on MSE, we repeat these experiments with one minor modification. Instead of simulating bridging for the entire period, we limit our bridging to only 10 years. Figures 4 and 5 summarize the results of these experiments. The trends visible in these figures are very similar to those shown in Figures 2 and 3. Specifically, it is clear that bridging brings about a substantive reduction in MSE. The magnitude of the reduction seems to be proportional to the amount of bridge coding.

[Figures 4, 5 about here]

The second criterion we use to evaluate the effects of bridging is the percentage of the latent ability scores z_{ct} which are covered by 95% of the model's posterior latent scores estimates (coverage). Examination of the Figures 6-7 reveals that regardless of the direction of bridging, even one bridge coder (for the entire period) increases the coverage rates to approximately 90%-100%. This rapid improvement is contrasted with the slow improvement we see in coverage when we limit our bridging to only 10 years (see Figures 8-9). This result has important implications for the V-Dem project's approach to recruiting bridge coders to solve the constant country problem. In particular, we are likely to benefit far more from recruiting a small number of bridge coders to code full time series across constant and varying countries than we are from asking many coders to rate limited cross-sections.

[Figures 6-9 about here]

Next we turn to direct comparisons between the model’s parameters and the generated data. We begin by inspecting the z_{ct} parameters. By examining this we hope to assess the degree to which lack of variation in latent scores affects model fit, and the extent to which bridging helps alleviate potential issues. Before performing these direct comparisons, we follow Bafumi et. al (2005) and rescale the model’s fitted values so that they have the same mean and standard deviation as the original latent scores. To present these comparisons systematically, we plot the generated country latent scores against the model’s z parameters in Figures 10 and 11. The first thing noticeable in Figure 10 is that the estimates for the constant countries (in red) cluster in two distinct areas. One cluster is located close to the 45 degree line, and one below it.¹⁵ In addition, the results of the baseline model (upper left panel in both figures) indicate that the model tends to over-predict the latent scores for the random country. This is a result of the relativity of the latent scales and our decision to mean-center estimates to aid in comparability; the rotation of the latent traits is arbitrary. Focusing on Figure 10, we see that the quantity of bridge coding has a strong effect on model fit. Bridging pulls both the over predicted and the under predicted (constant) values toward the 45 degree line. Much like in previous diagnostics, this process is gradual. In the simulation in which all five coders from the constant country also code the random one (lower-right panel), the estimates for the random county are almost entirely clustered on the 45 degree line (apart from a number of outliers at the extremes of the distribution). Under this scenario, the estimates from the constant country are also closer to the 45 degree line.

[Figures 10 and 11]

The results presented in Figure 11 match closely with those observed in Figure 10. However, closer inspection of our results reveals that when the bridge coding is done from the constant country to the random one, the fit of the model is improved. As can be seen in Table 1, the MSE for both random and constant countries are lower for constant-random

¹⁵This looks suspiciously like a multiple-modes problem. A similar pattern does not appear when we use variational approximation to estimate the O-IRT model. Here we use Hamiltonian Monte Carlo.

bridging than for random-constant bridging. This result makes a lot of sense because, prior to bridge coding the model knows a lot less about the thresholds of raters in the constant country than it does about those in the random country. In terms of bridge-rater recruitment strategy, we see that we stand to obtain greater benefit from asking “Swiss” coders to code cases with substantial variation than we do from recruiting coders with well-estimated threshold values to code constant cases.

[Table 1 about here]

We now turn to inspecting model fit with regards to the coders’ thresholds parameters $\gamma_{j,k}$. As in the previous examples, we compare patterns and quantity of bridging, how these affect the model’s predictions, and their relationship with the latent generated data. Figures 12 and 13 summarize these experiments. As always, the no-bridging scenarios yields estimates that are biased. Specifically, the threshold parameters of coders who only rate the random country are over-predicted, while those of coders who only code the constant country are under-predicted. As can be seen in Figure 12, increasing the degree of bridging from the constant to the random country reduces the bias, as more and more of the points fall closer to the 45 degree line. In addition, it is clear that even without a large degree of bridging, the lower thresholds of the constant coders are estimated with minimal bias.¹⁶ Overall, even with full bridging (i.e. when all coders from the constant country also code the random country), there is still bias in the estimates of coders’ thresholds. Examination of figure 13 highlights the extent of this issue. The figure shows that when the bridging pattern is from the random to the constant country, bridging has little effect on the bias with which these parameters are estimated.

In addition, the lower thresholds of the coders who are originally from the constant country (depicted at the upper panels of Figure 12 in blue dots and in the bottom panels in

¹⁶We simulate the constant country such that coders almost give it a perfect, high score. This minimal information is sufficient to estimate the location of the lowest threshold with little bias, but not of other thresholds.

red, as we simulate more and more bridge coding), are estimated with minimal bias when we simulate full bridging¹⁷

[Figures 12 and 13]

3.2 Thresholds Drawn from Different Distributions

We now turn our attention to situations where latent traits exhibit substantial variation within countries but coders across countries have significantly different thresholds. To simulate the situation where coders from different countries have different thresholds, we generate the coders' thresholds drawn from different distributions. For the "low threshold" coders, the means of their k thresholds, $\gamma_{\mu,k}^{low}$, are drawn from a $\mathcal{N}(-1, .5)$ distribution. These coders are able to discriminate between lower-valued latent scores, but lump high latent values into their top ordinal category. For the "high threshold" coders, the means of their k thresholds, $\gamma_{\mu,k}^{high}$, are drawn from a $\mathcal{N}(1, .5)$ distribution, subject to the ordering constraint. These coders are less able to discriminate low latent traits. This simulates a country where coders consistently tend to be more harsh on their evaluations of the country's level of political development. These two sets of coders rate two "random" countries, for which the latent scores are both drawn from a $\mathcal{N}(0, 1)$ distribution. The parameters of coders' discrimination are generated from the same distribution as in the previous simulation. We also follow the same data generating process to simulate coders' observed ordinal ratings for a hundred years of two countries.

In the baseline (unbridged) model, five "low threshold" coders rate only one country, while the five "high threshold" coders code the other country. Then we evaluate the effects of bridge-coding by simulating either "low threshold" or "high threshold" coders' ratings for the other country for the entire period. The MSE of these simulations are shown in Figure 14 and 15. Figure 14 and 15 demonstrate the effects of bridging coding done by "high threshold" and "low threshold" coders, respectively, on the MSE. The Figures suggest

¹⁷That is, a situation where all countries from country A also code country B.

that bridging, either done by the “high threshold” or “low threshold” coders, brings a rapid improvement for the “bridged country.” However, the improvement for bridge coders’ original country is not substantial. For example, when a “low threshold” coders does bridging coding for the other country, it provides information about where thresholds of those “high” coders are. In addition, the “high” coders’ are less able to distinguish between latent traits below their lowest threshold, and the bridging coding helps solve this issue. Similarly, the “low” coders cannot distinguish between latent traits above their highest threshold, but since none of the “high” coders does bridging coding for the “low” coders’ country, those high latent traits are still not accurately estimated.

[Figures 14, 15 about here]

Figures 16 and 17 more clearly illustrate the effects of bridging coding on the country-year estimates. The Figures show that in the baseline experiment, country-years coded by the “low” coders (black dots) tend to be overestimated; while country-years coded by the “high” coders (red dots) are likely to be underestimated. In addition, the high latent traits in the “low” coders’ country are in general poorly estimated, and vice versa. The estimates for both countries move toward the 45-degree line. However, for the bridging countries, the high latent traits in the “low” coders’ country and the low latent traits in the “high” coders’ country are still poorly estimated.

[Figures 16, 17 about here]

In sum, these experiments imply that, when coders in different countries have significantly different standards—when one set of coders is more strict than the other—we require bridges in both directions to obtain scale equivalence. It also demonstrates that the information provided by limited bridging does not filter through $\gamma_{r,k}$ and z_{ct} estimates as rapidly as one might intuitively expect. Of course these simulations are, in a sense, worst case, because the cross-national threshold differences that we examine here are severe. Nonetheless, because

we can not know how different coders thresholds are a priori, these simulations expose the potential severity of the bridging problem.

4 Identifying Poorly Bridged Cases 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 best mitigate this issue. The experiments with simulated data suggest that bridge rating helps solve the issues of “constant” countries and different thresholds of different groups of coders. When the quantity of bridge coding increases, the MSE between the country-year estimates and the simulated latent scores decreases. But our ability to evaluate bridging quality in these simulations rests on the fact that we know true latent values. To address these question with real data, we need techniques for identifying poorly bridged cases from observable quantities.

This is a difficult problem because the question of whether two cases are sufficiently bridged is, at the limit, unknowable. Nonetheless, while we may not be able to develop a test that establishes scale equivalence across sets of raters, we should be able to use the bridges that we do have to identify cases where scale equivalence is unlikely. In particular, our proposed approach uses in-sample posterior predictions from the model to test for poor model fit in bridge observations. Consider a bridge rater based in arbitrary country—we will call this bridge rater r_b — who provides a rating $y_{ctr_b} = k$ for a case, ct , in another country. Given model parameters, the posterior predicted probability of observing the bridge rating $y_{ctr_b} = k$ is

$$p(y_{ctr_b} = k | \boldsymbol{\gamma}, \mathbf{z}, \boldsymbol{\beta}) = \Phi(\gamma_{r_b, k} - z_{ct} \beta_{r_c}) - \Phi(\gamma_{r_b, k-1} - z_{ct} \beta_{r_c}).$$

A model that fits the data well should place relatively high probability on the outcome, k , that we actually observe. More importantly, all else equal,¹⁸ a model that exhibits scale equivalence across countries should do a similarly good job of predicting ratings by bridge

¹⁸Rater reliability, in particular.

raters as it does predicting ratings by in-country raters. On the other hand, if scales vary across countries, then the model should expect predict behavior by r_b that is off the mark, and it should do a worse job of predicting r_b 's ratings than it does predicting the ratings of in-country raters of the bridged case. Scaling up, if we find that our posterior predictive capability for bridge ratings substantially under-performs our ability to predict in-country ratings for bridged cases, then we should be suspicious that the target country's scale is out of whack with those in the bridge raters' countries.

Levering this logic, we calculate the posterior probability based on the point estimates of γ_{r_b} , z_{ct} , and β_{r_c} for each bridge codings,¹⁹ and take the averages across all years and all bridge coders within each bridged country to generate the country level probability $p(y_{c_b} = k)$. Focusing only on bridged observations, we also calculate the same posterior predicted probability of observing y_{ctr_c} for each in-country rater r_c , and then generate the country level average for each bridged country, $p(y_{c_c} = k)$. These average scores tell us how well the model fits bridge raters relative to in-country raters for bridged cases. By focusing only on bridged cases, we control for the degree of difficulty inherent in predicting the cases that happen to be bridged. For a bridged country c , if these two values are close to each other, the model fit is similar across bridge and local raters for this country. Thus we have little evidence of scale incompatibility. On the other hand, if the model fits bridge raters poorly, relative to in-country coders, then we know that scale equivalence is unlikely.

It is important to note that such comparison cannot establish scale identification across countries. Rather a small difference between the posterior probabilities indicates only that there is little evidence that the scale of measurement in country c is different than the scales in the countries that supplied bridge raters for c . On the other hand, if the difference is large, then the model is substantially worse at predicting how bridge raters will judge country c than it at predicting how local rater behavior. There are a number of possible reasons for such a finding. For instance, bridge coders might be less discriminating than

¹⁹We also calculate a posterior distribution around the probability based on each draw from the model posterior.

their in-country counterparts. But substantial differences between how well the model fits bridge codings and local codings of bridged cases can help researchers identify cases that are unlikely to be globally identified, and thus require more bridge coding.

[Figure 18 about here]

For the experiments on two types of countries described in section 3.1, we calculated these average posterior probabilities. Figure 18 shows the results and compares the scores of in-country and bridge coders when there are different numbers of bridge coders from the “random” country rate the “constant” country. As expected, the posterior probabilities of bridge ratings are lower than those of in-country ratings. In addition, the differences between them decrease as more coders from the “random” country provide ratings for the “constant” country. In the baseline model where there is no bridging, the posterior probabilities of the in-country coders are higher than those of bridge coders by around .15. When all coders from the “random” country code the “constant” country for all the years, the average difference between them decrease to .03.

[Figure 19 about here]

We also calculate the posterior probabilities for experiments in which coders from the “constant” country do bridging coding for the “random” country. The results are shown in Figure 19. The posterior probabilities of the in-country and those of bridge coders do become closer when there are more bridge coders. However, to our surprise, the posterior probabilities of the bridge ratings for the “random” country are consistently higher than those of in-country ratings. The probabilities of the bridge ratings *decrease* with the number of bridge coders, and therefore their probabilities become closer.

This result calls into the question the validity of this PPD-based test for bridging failure. We are currently investigating other tools for identifying scale equivalence problems in real data.

5 Efficiently Selecting Bridges

The simulations in section 3 illustrate general patterns in bridge rating efficacy. Ideally, we would ask a significant number of country-coders to provide bridge ratings for long time periods in a second country. But, in projects like V-Dem, where coding demands significant investment from experts, we would like to use patterns in the ratings that we do have to identify potential bridges that are likely to give us the most bang for the buck. Our approach to this problem is motivated by the literature on computerized adaptive testing (CAT).²⁰

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—conserving resources. Here, students are analogous to country-years and questions take the place of raters. One generally assumes that question (rater) parameters—each $\gamma_{r,k}$ and β_r —are known and iteratively chooses questions, based on their estimated parameters that minimize the expected posterior variance in the estimate of the student’s latent ability. First, 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 generally involves presenting a student with a moderately difficult question to begin with, presenting a harder (easier) question if the student does well (poorly) on the first question, and repeating 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

$$\operatorname{argmin}_q \left[\sum_{k \in 1}^K p_q(y_q^{i+1} = k | \mathbf{y}^i) \operatorname{Var}(z | \mathbf{y}^i, y_q^{i+1} = k) \right] : q \in \mathbf{Q}_i \quad (4)$$

²⁰See Montgomery & Cutler (2013) for an introduction to this literature aimed at survey researchers in political science. Choi & Swartz (2009) provide an overview of CAT item selection methods for ordinal items.

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 , $\text{Var}(z | \mathbf{y}^i, y_q^{i+1} = k)$ 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 i th step of the process (Choi & Swartz 2009).

5.1 Computerized Adaptive Bridging

Our problem differs from the traditional CAT setting in a number of ways. Most importantly, we cannot assume that our current parameter estimates are valid. Our primary goal here is not to reduce uncertainty, but rather to find and eliminate bias. If latent trait estimates are biased at step i , then a bridging strategy based on MEPV has the potential to erroneously increase our certainty in both biased threshold estimates and biased latent traits. A bridge-selection strategy based on MEPV chooses bridges that are most likely to reinforce existing model estimates. Nonetheless, the logic of CAT suggests a strategy for efficient bias detection and correction. When a bridging problem exists, good bridges will substantially alter parameter estimates, particularly z_{ct} and $\gamma_{r,k}$ values. This is because good bridges are cases in which a bridge rater behaves in a way that is highly inconsistent with the model’s expectations. Therefore, focusing on latent traits, we propose selecting bridges that have the highest potential to alter existing estimates. Formally, at each step in the algorithm, we select the bridge that meets the condition

$$\operatorname{argmax}_{\{c,t,r\}} \left[\frac{1}{K} \sum_{k=1}^K \left(\frac{1}{|C| \times |T|} \sum_{c,t \in C \times T} [E(z_{ct} | \mathbf{y}^i, y_{ctr}^{i+1} = k) - E(z_{ct} | \mathbf{y}^i)]^2 \right) \right] : \{c, t, r\} \in \mathbf{B}_i \quad (5)$$

where \mathbf{B}_i is the set of potential remaining bridges at step i . This approach iterates over potential bridges and chooses the bridge that maximizes the mean square error between current and potential estimates, averaged across potential bridge rater responses. We call this the MAM, or maximum average MSE, algorithm. This approach ignores the expected

probability that bridging raters will choose each category, k . This is a potential feature because, when the model lacks cross-national scale equivalence, predicted response probabilities will be misleading. Nonetheless, as the model approaches scale equivalence, weighting expected MSE by the probability of observing each particular bridge rating will guide the algorithm towards more informative bridging opportunities. Thus, we also propose an alternative algorithm—the maximum expected MSE, or MEM—which iteratively selects bridges meeting the following condition

$$\operatorname{argmax}_{\{c,t,r\}} \left[\frac{1}{K} \sum_{k=1}^K p_{ctr}(y_{ctr}^{i+1} = k | \mathbf{y}^i) \left(\frac{1}{|C| \times |T|} \sum_{c,t \in C \times T} [E(z_{ct} | \mathbf{y}^i, y_{ctr}^{i+1} = k) - E(z_{ct} | \mathbf{y}^i)]^2 \right) \right] : \{c, t, r\} \in \mathbf{B}_i. \quad (6)$$

In principle, we should be able to run either of these algorithms until the observed MSE between parameters from the pre- and post-bridge model fits falls below a desired threshold. Moreover, we can easily adapt these algorithms 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 rate 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 is unlikely to be practical to recruit bridges one at a time in most cross-national expert survey projects. Thus, rather than attempting to recruit only the single bridge that maximizes equations 5 or 6, 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 proposed algorithms for efficient bridge search are very computationally expensive because they require 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 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 pa-

parameter, making numerical integration a practical approach (Montgomery & Cutler 2013); here we need to estimate full multi-parameter models, and solve high-dimensional integrals, 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).²¹ Appendix A provides details.

5.2 Evaluating the MAM Algorithm

Figure 20 depicts the results of a simulation that compares the MAM algorithm to a random search for bridges. Specifically, we simulated an initial unbridged dataset just as we did when examining the constant coder problem in section 3.1.²² Then we simulated two search processes. Each process sequentially added 100 bridges to the unbridged dataset. After each bridge was added, we calculated the mean square error between the “true” distribution of latent traits and the estimates provided by the O-IRT model. In the first search process, we chose bridges at random. In the second, we used the MAM algorithm to sequentially select bridges that had the maximum average potential to alter latent trait estimates. As figure 20 clearly shows, the MAM algorithm chooses bridges that reduce the MSE substantially more rapidly than is possible through random selection. Indeed, we obtain an impressive drop in MSE with fewer than 25 bridge codes, approaching that of a fully bridged model (i.e. every rater rates every case). Yet clearly, the MAM process exhibits sharp discontinuities in MSE. This is because, while the algorithm identifies bridges with the potential to “surprise” the model, the rater actually has to do something unexpected for the model to substantially recalibrate. It seems that potentially surprising bridges are also likely to be somewhat less informative than randomly selected bridges, when the rater does not behave in an unexpected

²¹Variational approximation techniques were developed by computer scientists 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 (2010) provides an introduction using examples drawn from political science.

²²Because the MAM algorithm is computationally expensive, we limit our time-spans to 50 years in this simulation.

way. This aspect of the algorithm is likely to stymie the search for a reliable stopping rule, and may be related to the failure of PPD comparisons to effectively diagnose bridging issues. Nonetheless, the algorithm shows promise. Future iterations of this paper will investigate how the MEM algorithm fares in this regard. It would also be interesting to see how the algorithm fares when given tasks such as choosing full-time-series bridge-raters, or operating under potential bridge constraints.

6 Conclusion

This paper provides perhaps the first systematic analysis of *bridging* in the context of item response models for cross-national expert surveys. We believe this to be an issue that is applicable to a large number of studies, which build upon the ability of country experts to provide detailed information about latent political concepts. While country experts' knowledge is a rich source of data, using this source brings about serious challenges in terms of cross-rater comparability, which, thus far, have not been thoroughly addressed in the literature. Furthermore, while the question has been more extensively examined in the context of measuring ideological common spaces, we lack general tools for effectively assessing bridging quality in latent variable models. Thus, exploring how patterns of bridging affect scale comparability in the sorts of models we examine here has implications for a wide array of work in the discipline.

Of course, our current study is motivated by the data collected as part of the *Varieties of Democracies* project. Therefore, we have focused on patterns of bridging failure that are relevant to V-Dem. We address the issue of bridging low-variance, 'constant' countries, by performing a number of Monte Carlo simulations in which we gradually increase the degree of bridging to, and from, these countries. Our results demonstrate that bridging has a large impact on the model's parameters. First, the results indicate that bridging brings about a rapid decrease in the mean square error of the latent ability scores, and an increase in

credible interval coverage. In our simulated data, even one bridge coder (a coder who rates both countries for the entire 100 year period), brings about a reduction of about 70% in MSE, and a 30%-50% increase in coverage rates. However, it is important to note that the pattern of bridging has consequences. The results show that recruiting a limited number of bridge coders who rate a large period in the bridged country is a much better way of reducing bias than obtaining a large number of coders who only rate a small number of years.

In conjunction with examining MSE and coverage, a direct comparison of the model's estimated parameters with the data we generated yields an additional important finding. Specifically, when it comes to estimating countries' latent ability scores (z_{ct}), bridging from a constant country to a country with substantive variation provides a larger reduction in model bias. When we inspect the effects of bridge coding on estimation of thresholds ($\gamma_{j,k}$ parameters), it becomes evident that a constant-random bridging pattern is a much better way of reducing bias and identifying coders' thresholds than a random-constant bridging pattern.

Next, we tackle the issue of cross-rater thresholds' comparability. We know that in many cases, raters vary greatly in their thresholds. Some coders might be described as lenient, i.e., they have very low thresholds for democratic indicators, while others may be tough, and tend to high higher standards in terms of their rating. We address this issue by drawing thresholds from two distributions with different means. Simulating bridging for the two types of coders shows that bridging brings about an improvement in terms of MSE and of coverage, but that this improvement is limited. Specifically, we gain improvement in estimating z_{ct} parameters from the country from which we simulate the bridging, but not for the target country. Overall, our experiments show how serious the varying threshold problem can be, and that a very large amount of bridging (possibly from both directions) is necessary in order to overcome bias that results from this issue.

The severity of bridging issues requires a technique that would help in evaluating bridging quality. To do this, we utilize the posterior predictive probability (PPD) of having a given

rating for in-country coders ($p(y_{c_c} = k)$), and compare this (averaged) probability with the probability of a given rating for bridge-coders ($p(y_{c_b} = k)$). Our goal was to examine the degree to which rates and patterns of bridge coding brings about a change in the comparability of bridge coders' correct prediction probabilities with that of in-country coders. The results obtained from this technique indicate that a different approach might be more useful. Specifically, when we simulate bridge coding from a high-variance country to a constant one, we see that the increasing bridge-coding leads to higher similarity between in-country and bridge coders' prediction accuracy. This finding is in line with our theoretical expectations. However, when we simulate bridging from a constant to a high-variance country, the results do not match our expectations. First, we find that bridge coders from the constant country do a better job at predicting the raw ratings of the target country than in-country coders. In addition, increasing the amount of bridging leads to a decrease in the accuracy of in country coders.

Finally, we propose and evaluate an algorithm for efficiently choosing bridge raters and demonstrate promising results. Overall, the paper clearly shows that bridging (or lack thereof) matters. Ignoring this issues leads to misleading inferences, both in terms of latent ability estimates, and in terms of thresholds (difficulty) parameters. Moreover, achieving cross-rater comparability is a complicated task. Our analysis demonstrates that a large degree of bridging is necessary to bring about an improvement in biases that result from low variation in coders' rating, and from their varying conception of gradation in the latent concept that is being measured.

7 Figures

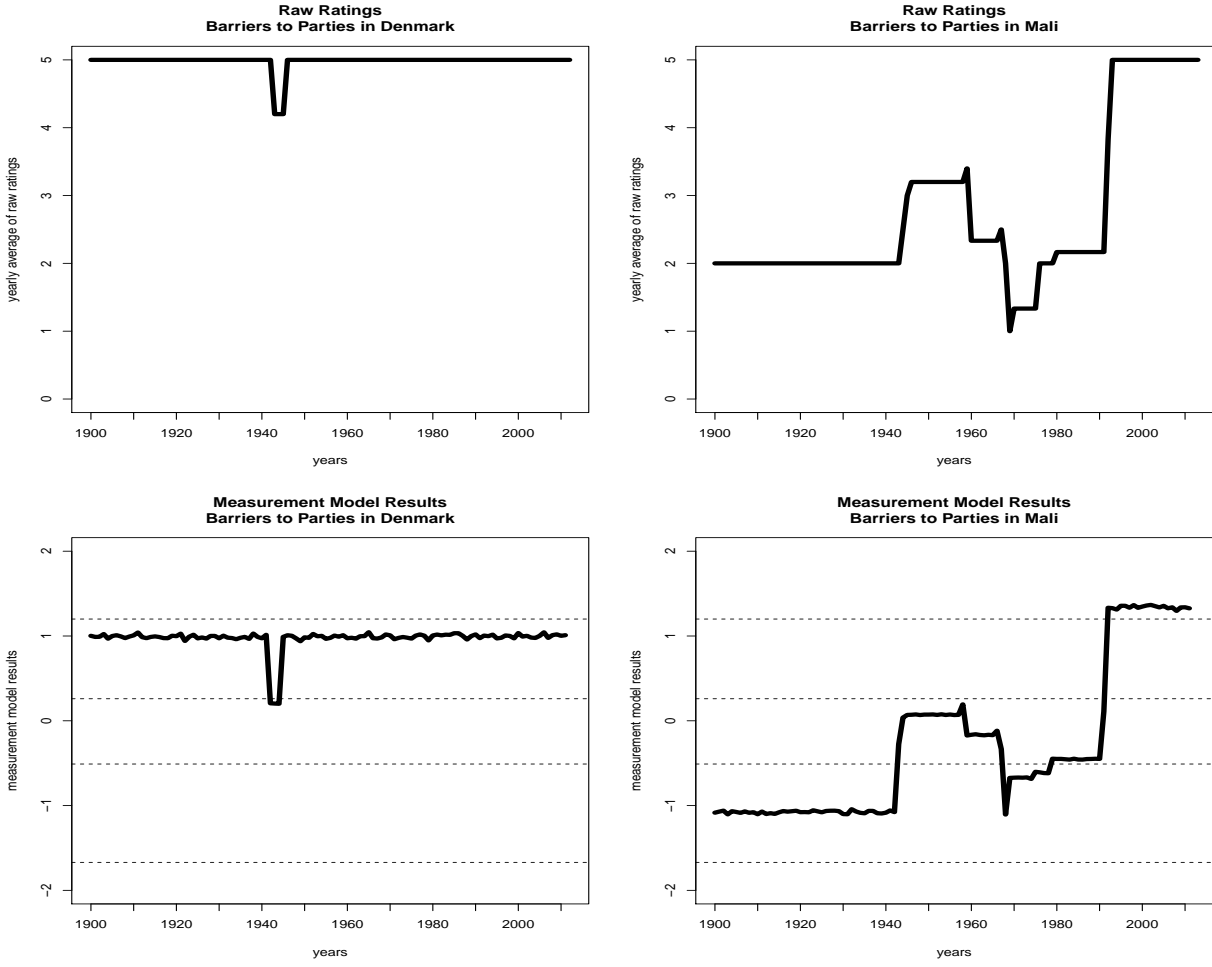
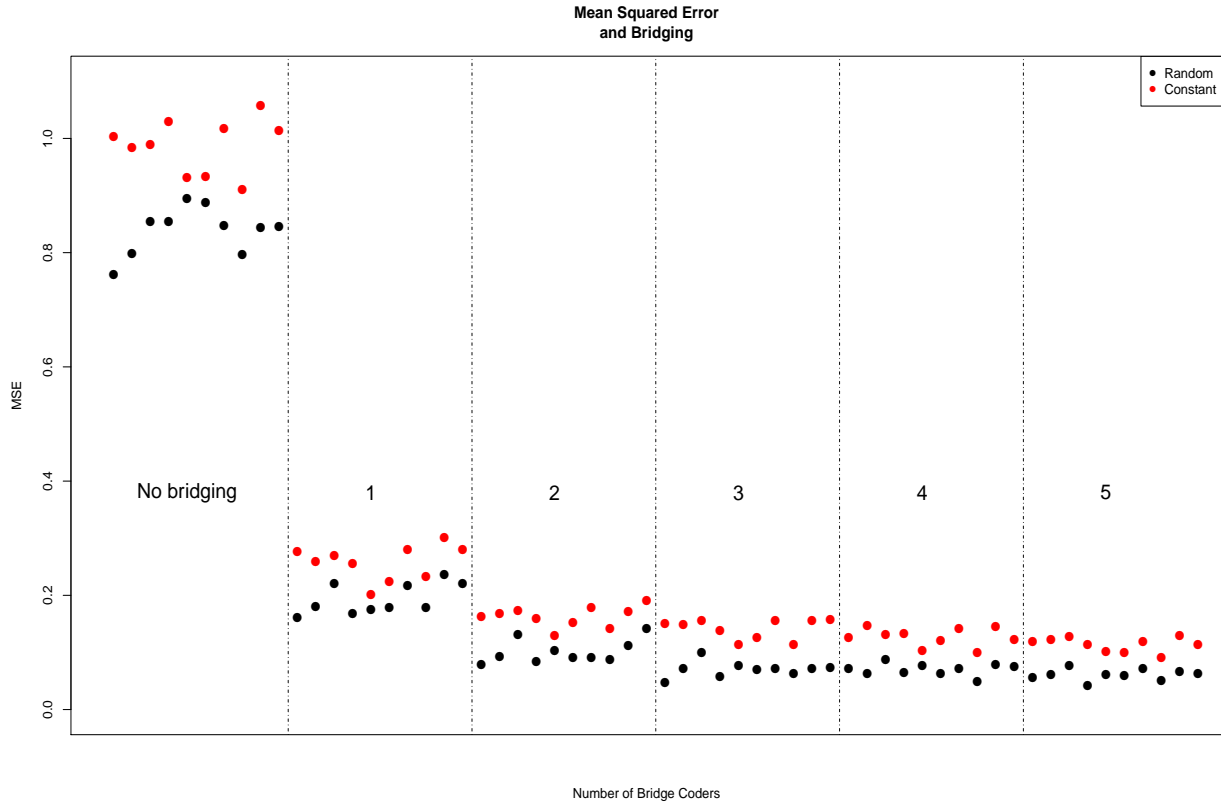


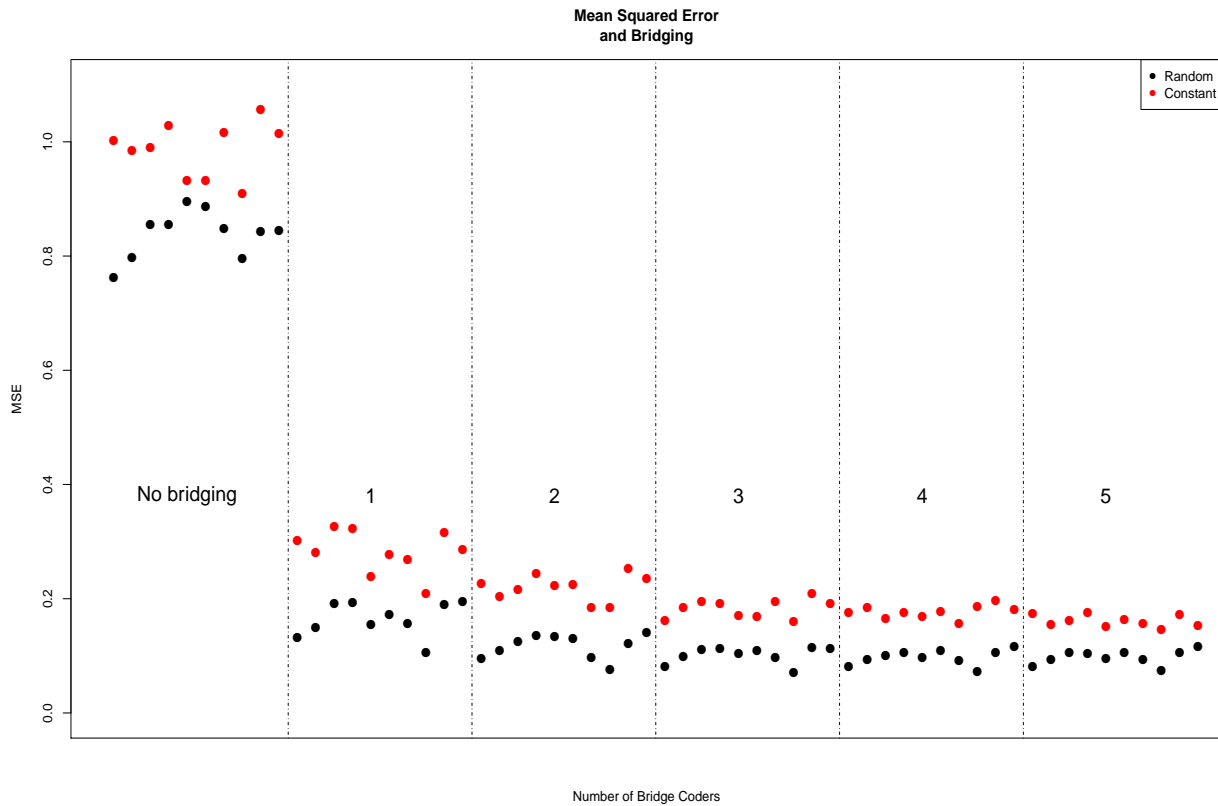
Figure 1: Raw ratings and measurement model results for *barriers to parties* indicator.



Note: The figure shows the effects of bridging from a constant to a random country. Bridge coding is done for the entire time period (100 years). The vertical lines divide the figure into the bridging patterns, from a no-bridging to full-bridging (i.e. all coders from the constant country also code the random country). In each section of the figure, points represent the MSE for 10 relevant simulations.

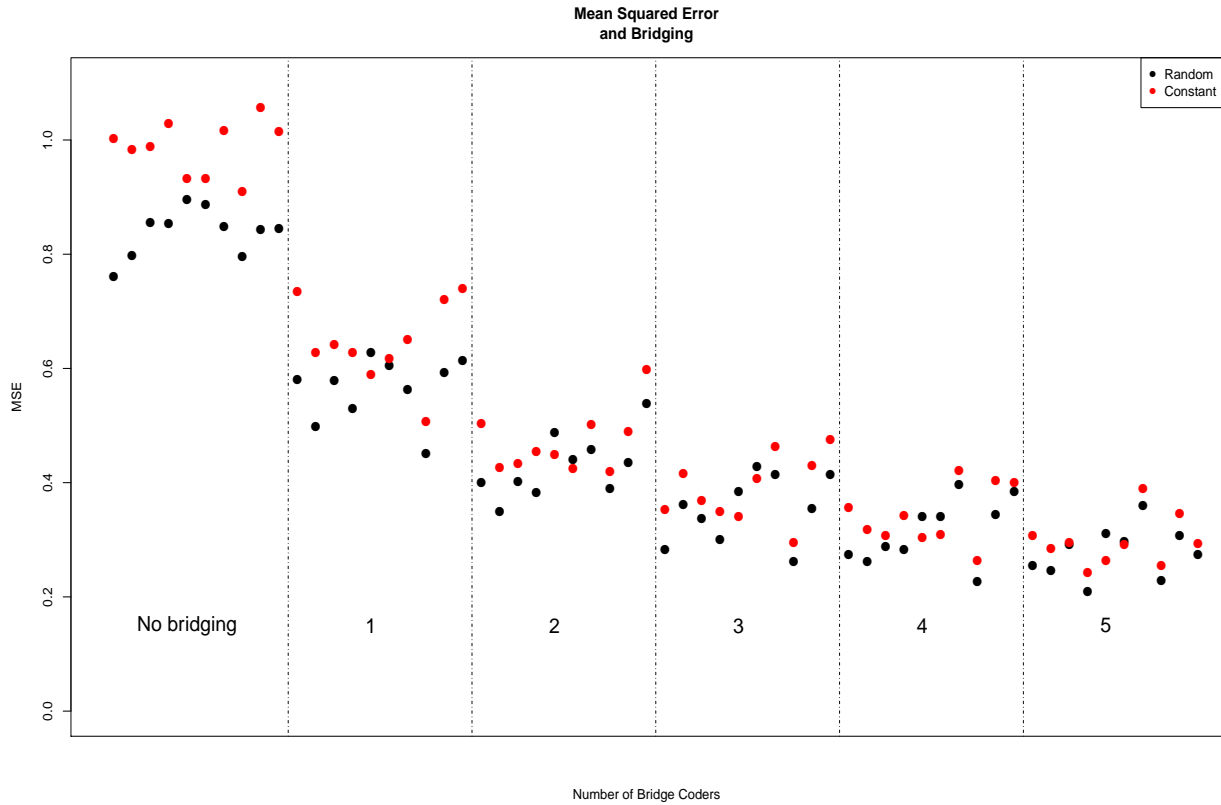
Figure 2: Evaluation of bridging from constant to random countries—mean square error of latent scores

similar, but not as strong as seen in figure 16. The bias for the random coders' thresholds is big, especially for the bridge coders.



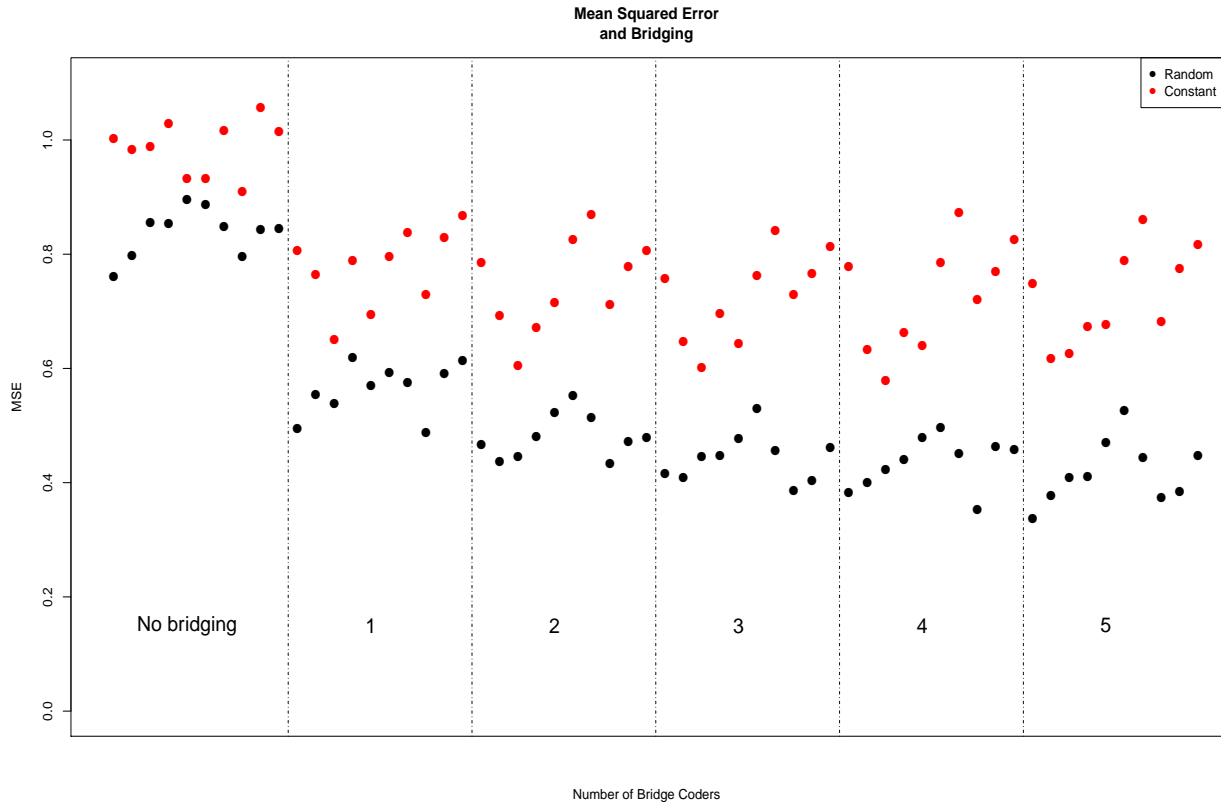
Note: The figure shows the effects of bridging from a random to a constant country. Bridge coding is done for the entire time period (100 years). The vertical lines divide the figure into the bridging patterns, from a no-bridging to full-bridging (i.e. all coders from the constant country also code the random country). In each section of the figure, points represent the MSE for 10 relevant simulations.

Figure 3: Evaluation of bridging from random to constant countries—mean square error of latent scores



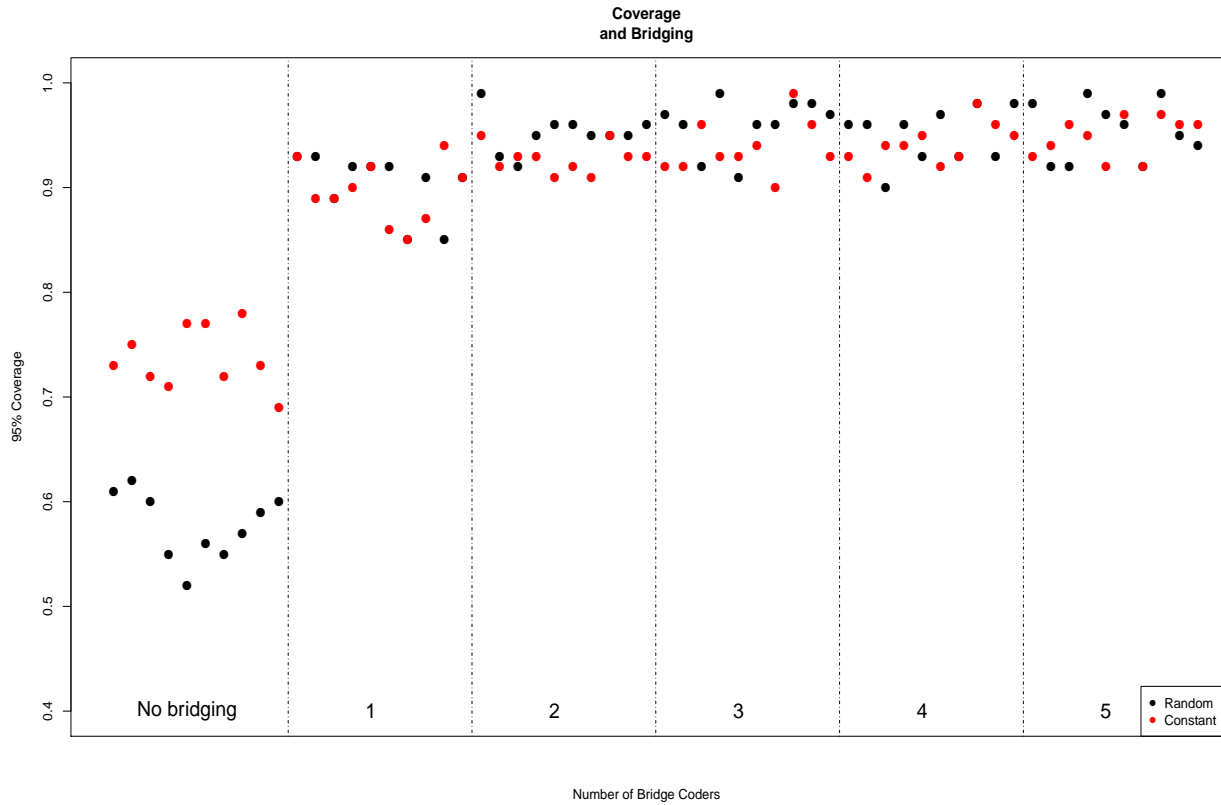
Note: The figure shows the effects of bridging from a constant to a random country. Bridge coding is done for 10 years. The vertical lines divide the figure into the bridging patterns, from a no-bridging to full-bridging (i.e. all coders from the constant country also code the random country). In each section of the figure, points represent the MSE for 10 relevant simulations.

Figure 4: Evaluation of bridging from constant to random countries (10 years of bridging)—mean square error of latent scores



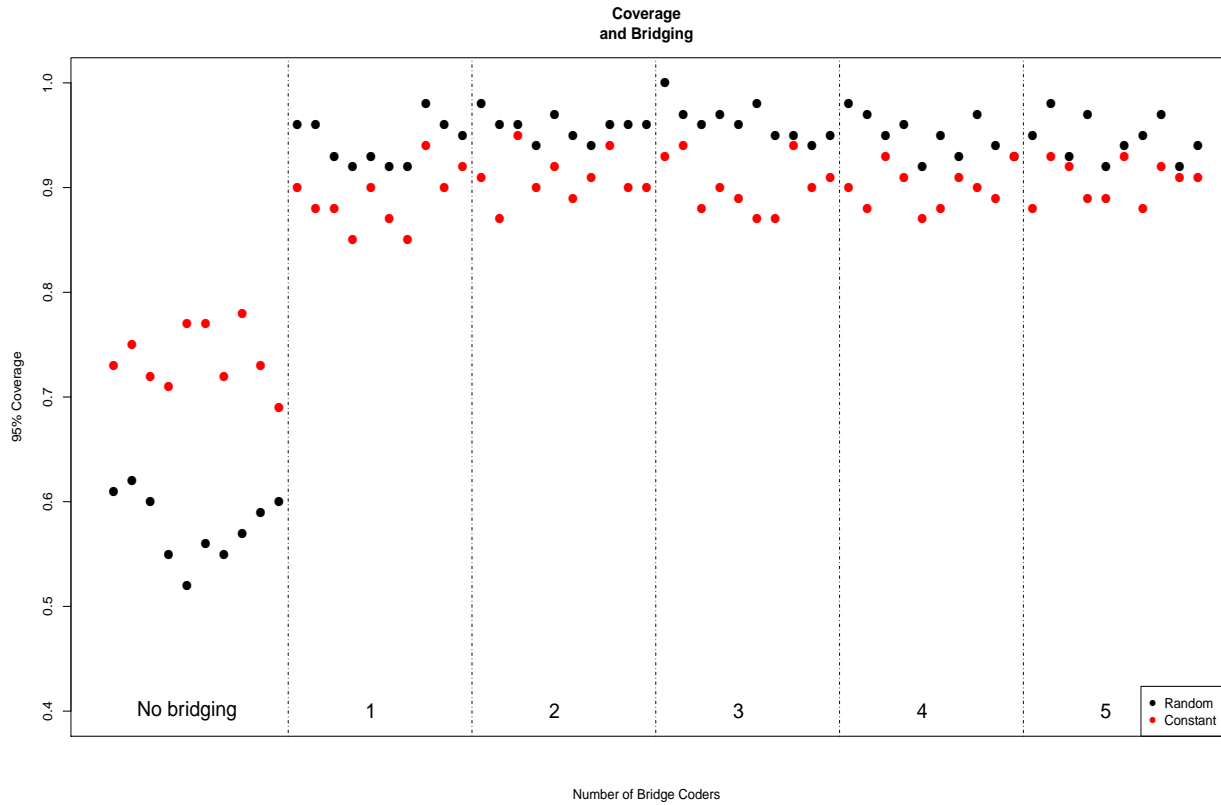
Note: The figure shows the effects of bridging from a random to a constant country. Bridge coding is done for 10 years. The vertical lines divide the figure into the bridging patterns, from a no-bridging to full-bridging (i.e. all coders from the constant country also code the random country). In each section of the figure, points represent the MSE for 10 relevant simulations.

Figure 5: Evaluation of bridging from random to constant countries (10 years of bridging)—mean square error of latent scores



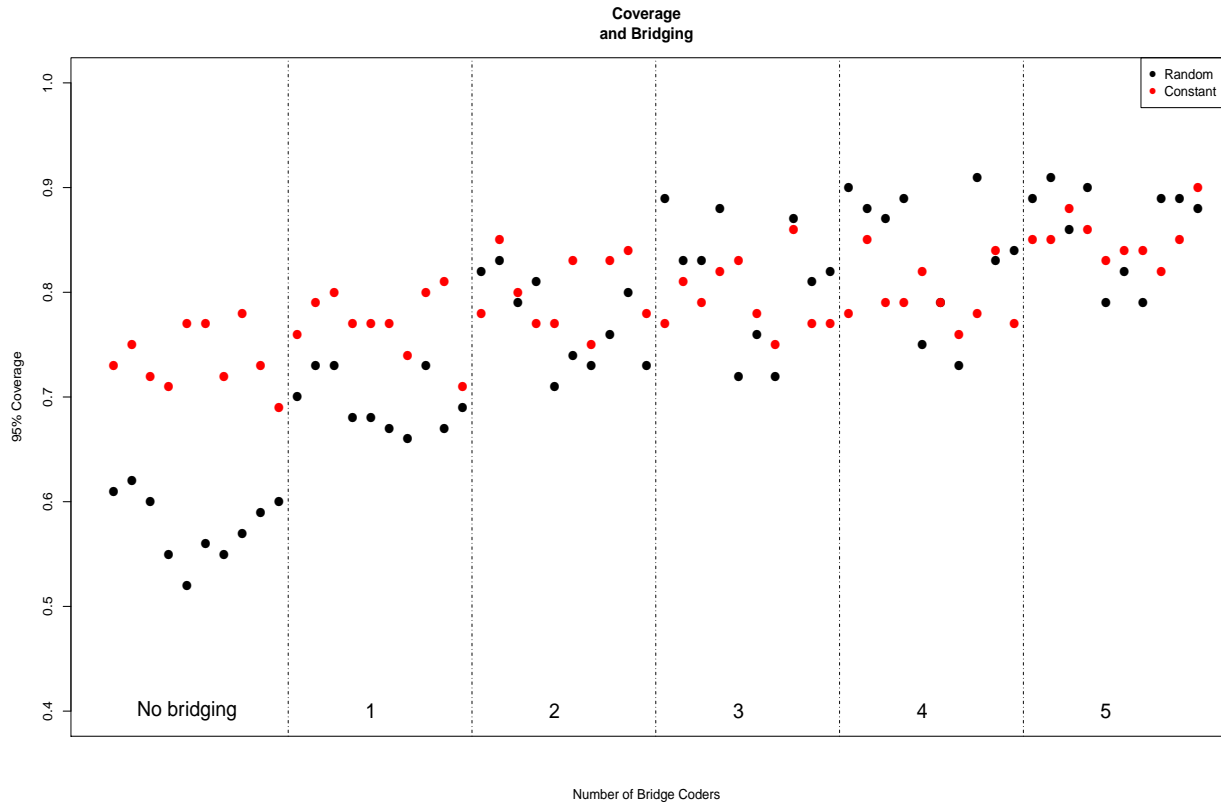
Note: The figure shows the effects of bridging from a constant to a random country. Bridge coding is done for the entire time period (100 years). The vertical lines divide the figure into the bridging patterns, from a no-bridging to full-bridging (i.e. all coders from the constant country also code the random country). In each section of the figure, points represent the coverage for 10 relevant simulations.

Figure 6: Evaluation of bridging from constant to random countries–95% coverage



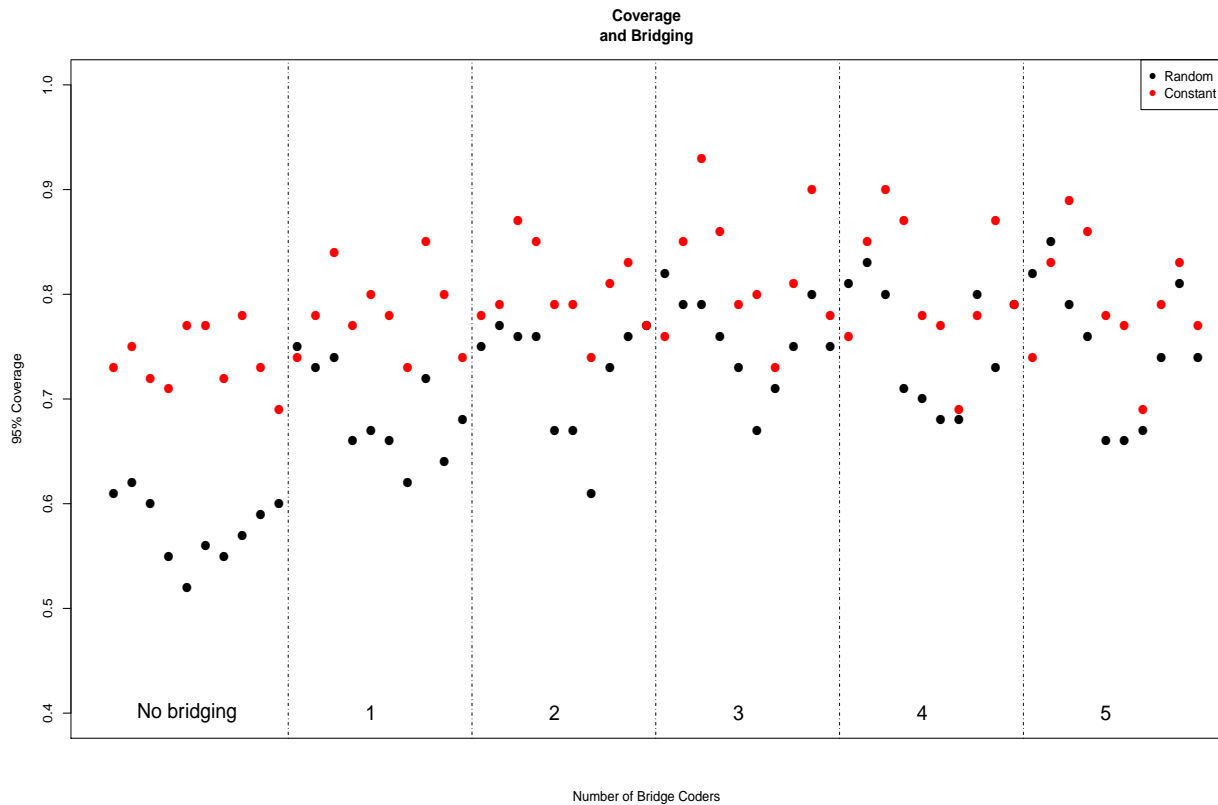
Note: The figure shows the effects of bridging from a constant to a random country. Bridge coding is done for the entire time period (100 years). The vertical lines divide the figure into the bridging patterns, from a no-bridging to full-bridging (i.e. all coders from the constant country also code the random country). In each section of the figure, points represent the coverage for 10 relevant simulations.

Figure 7: Evaluation of bridging from random to constant countries–95% coverage



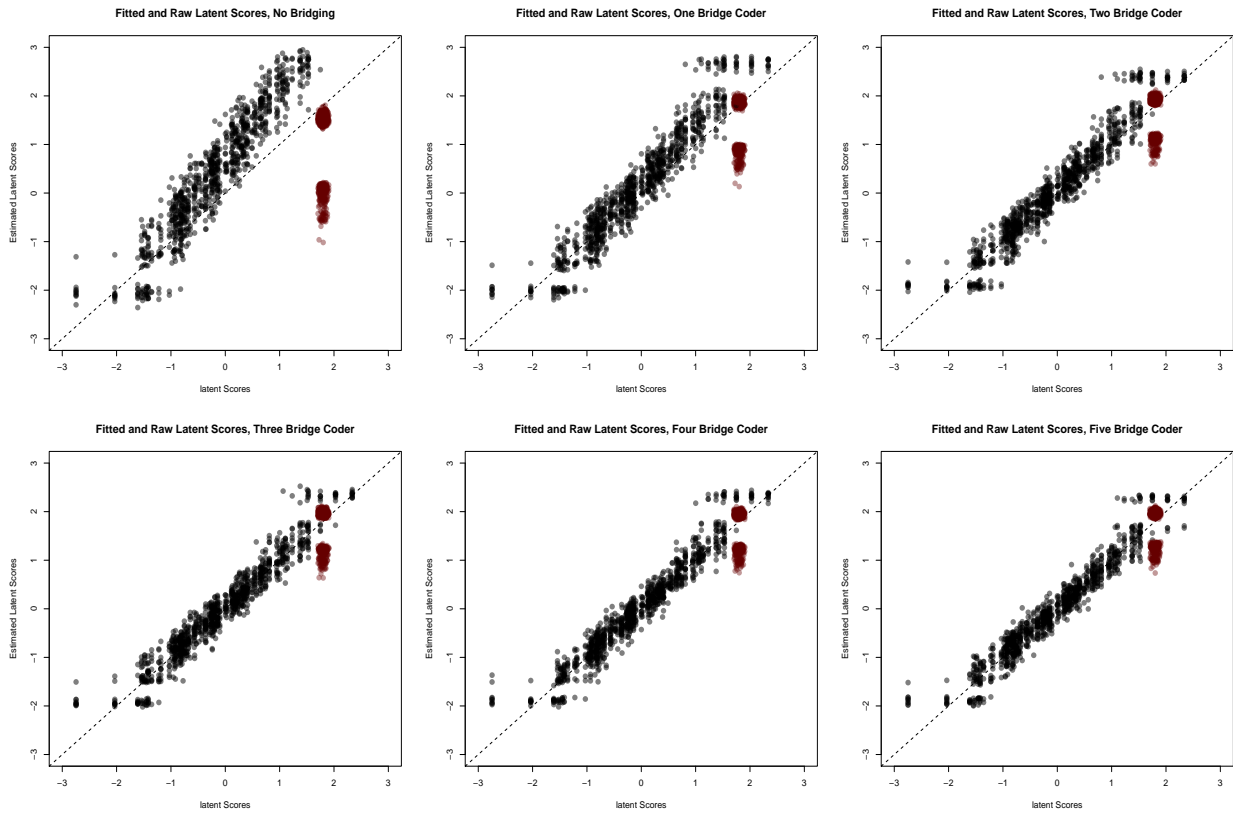
Note: The figure shows the effects of bridging from a constant to a random country. Bridge coding is done for a 10 year period. The vertical lines divide the figure into the bridging patterns, from a no-bridging to full-bridging (i.e. all coders from the constant country also code the random country). In each section of the figure, points represent the coverage for 10 relevant simulations.

Figure 8: Evaluation of bridging from random to constant countries–95% coverage



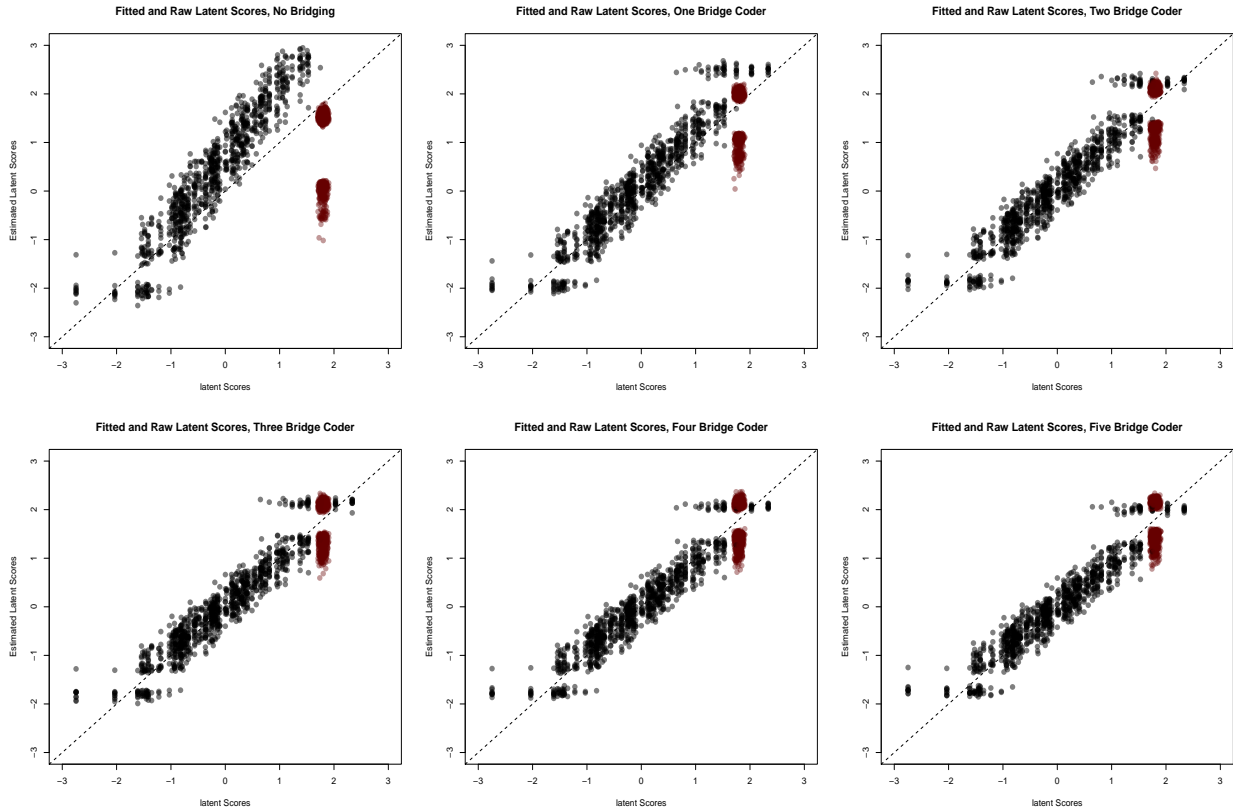
Note: The figure shows the effects of bridging from a random to a constant country. Bridge coding is done for a 10 year period. The vertical lines divide the figure into the bridging patterns, from a no-bridging to full-bridging (i.e. all coders from the constant country also code the random country). In each section of the figure, points represent the coverage for 10 relevant simulations.

Figure 9: Evaluation of bridging from random to constant countries—95% coverage



Note: The figure shows the effects of bridging from a constant to a random country. The upper left panel presents the results for the baseline mode (no bridging). Points depict the results of per simulations per regime (no bridging, one bridge coder, two bridge coders, etc.)

Figure 10: Fitted and real latent score estimates. Bridging is from constant to random countries



Note: The figure shows the effects of bridging from a random to a constant country. The upper left panel presents the results for the baseline mode (no bridging). Points depict the results of per simulations per regime (no bridging, one bridge coder, two bridge coders, etc.)

Figure 11: Fitted and real latent score estimates. Bridging is from random to constant countries

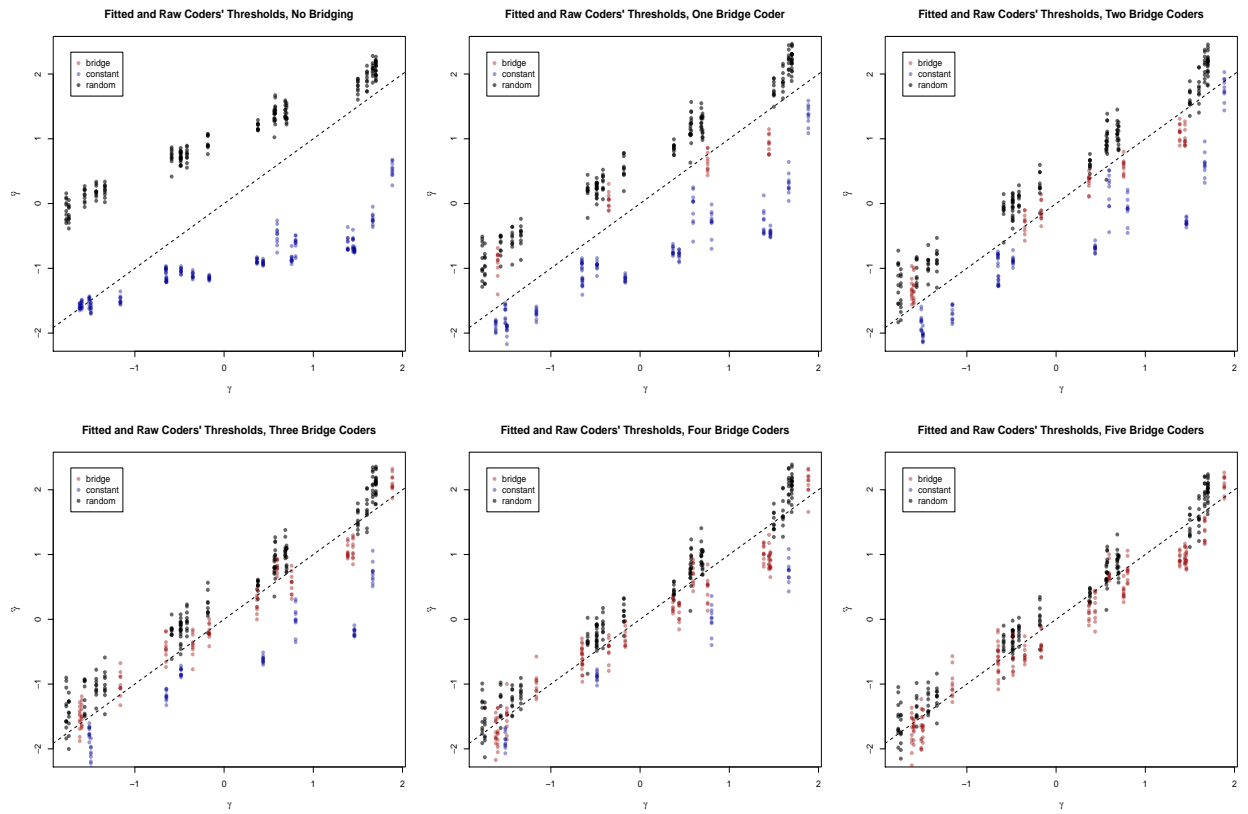


Figure 12: Fitted and real thresholds' estimates. Bridging is from constant to random countries.

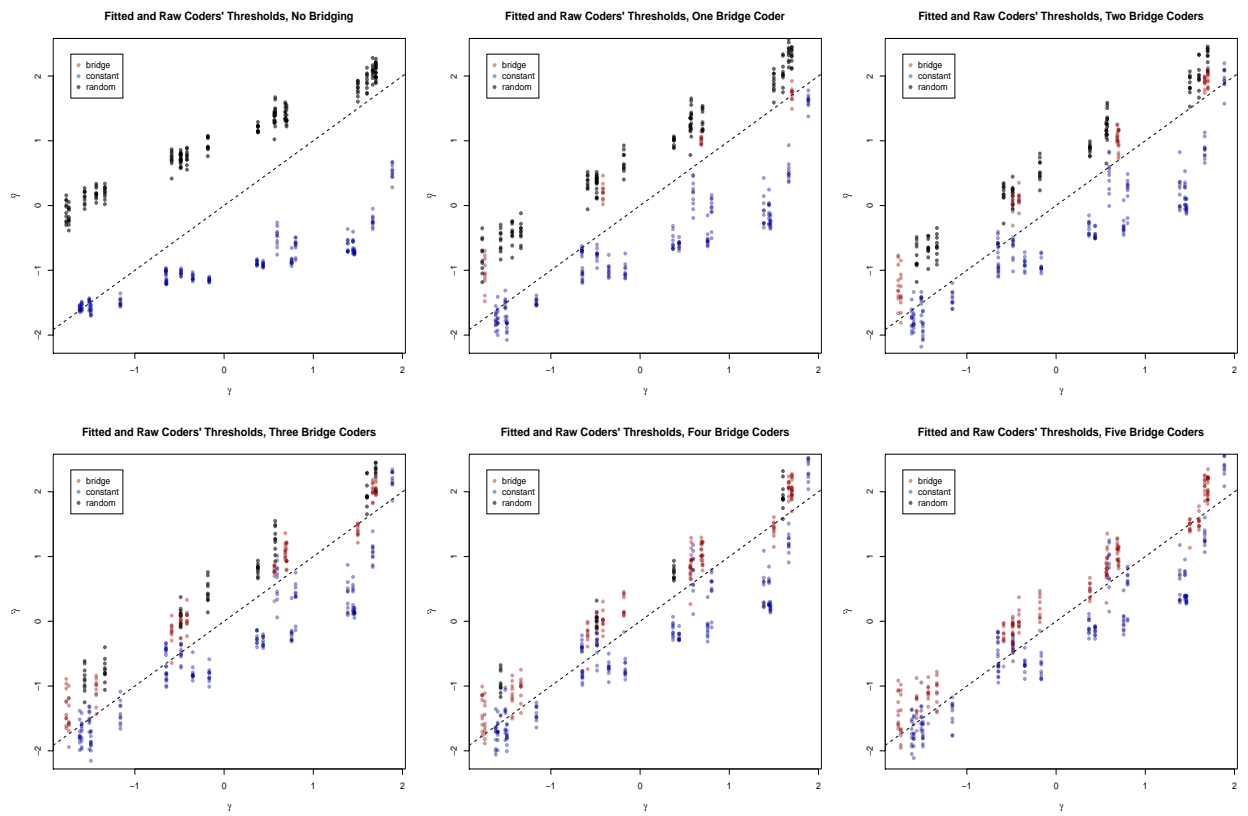


Figure 13: Fitted and real thresholds' estimates. Bridging is from random to constant countries.

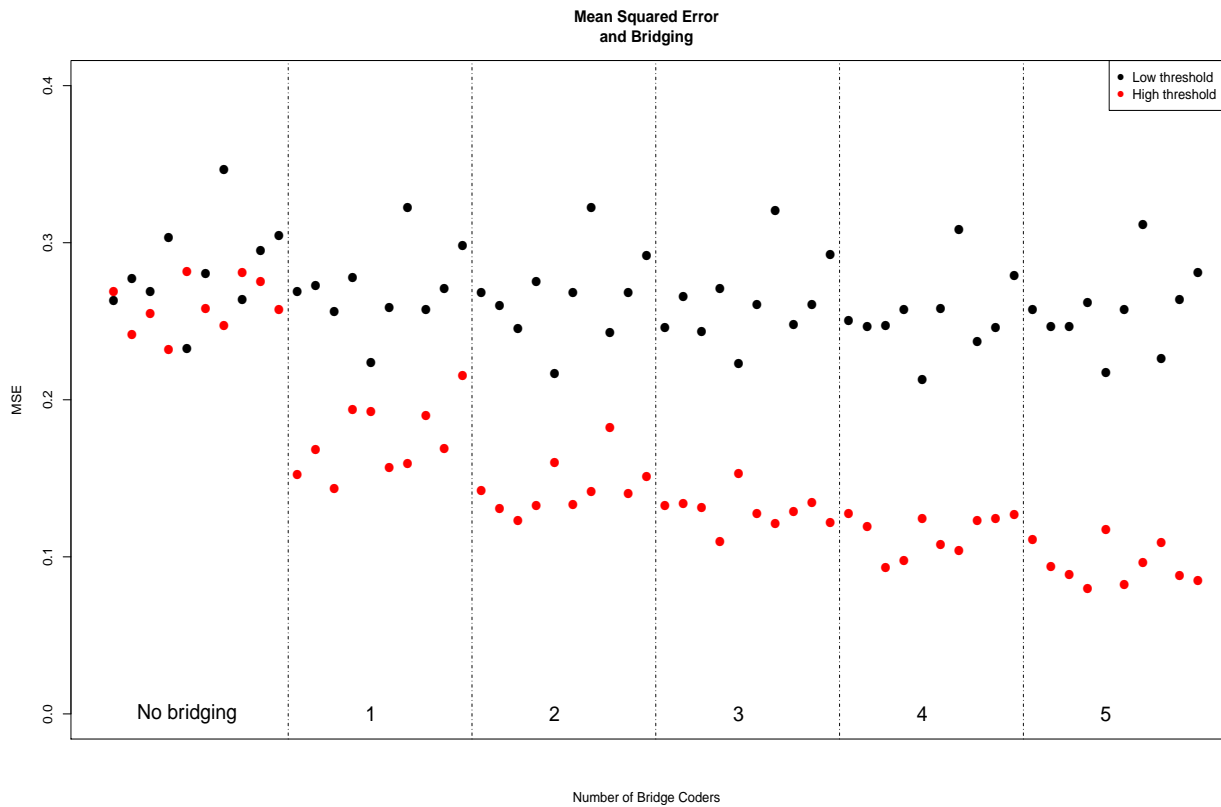


Figure 14: Evaluation of bridging from a country with low coders' thresholds to a country with high coders' thresholds, mean square error of latent scores.

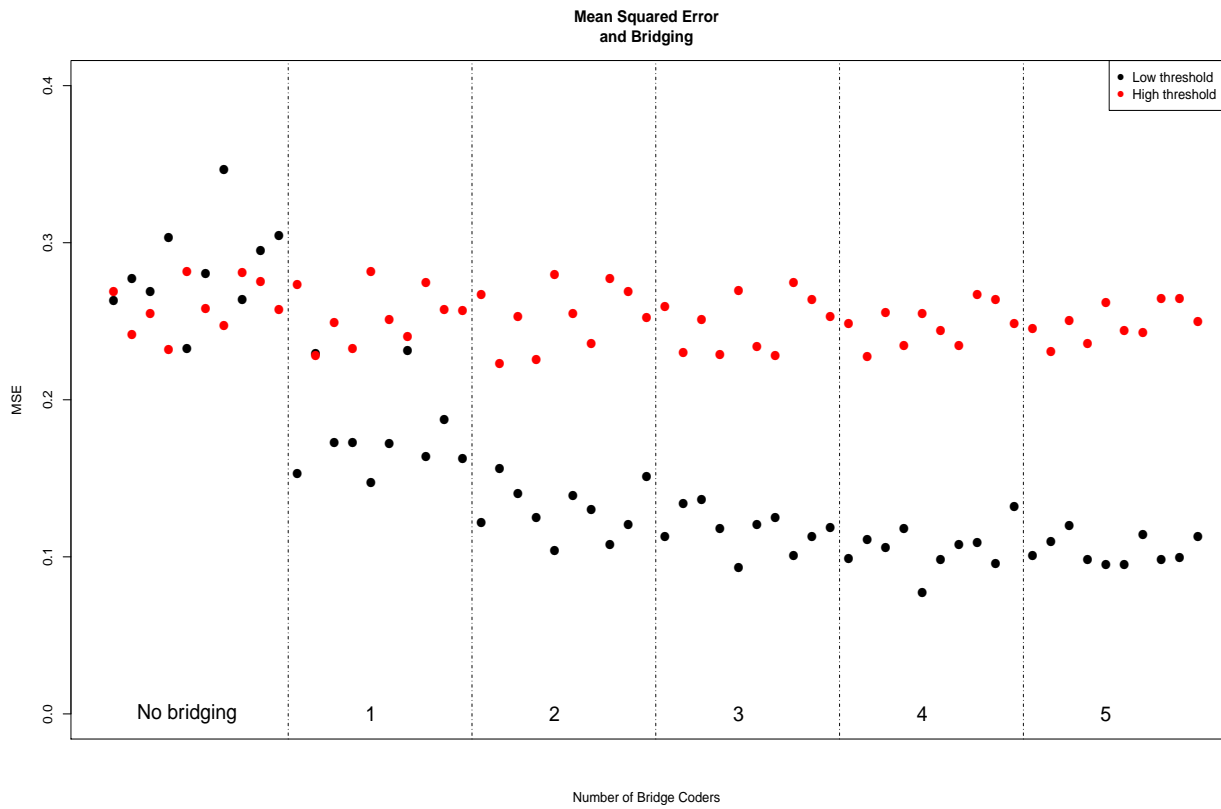


Figure 15: Evaluation of bridging from a country with high coders' thresholds to a country with low coders' thresholds, mean square error of latent scores.

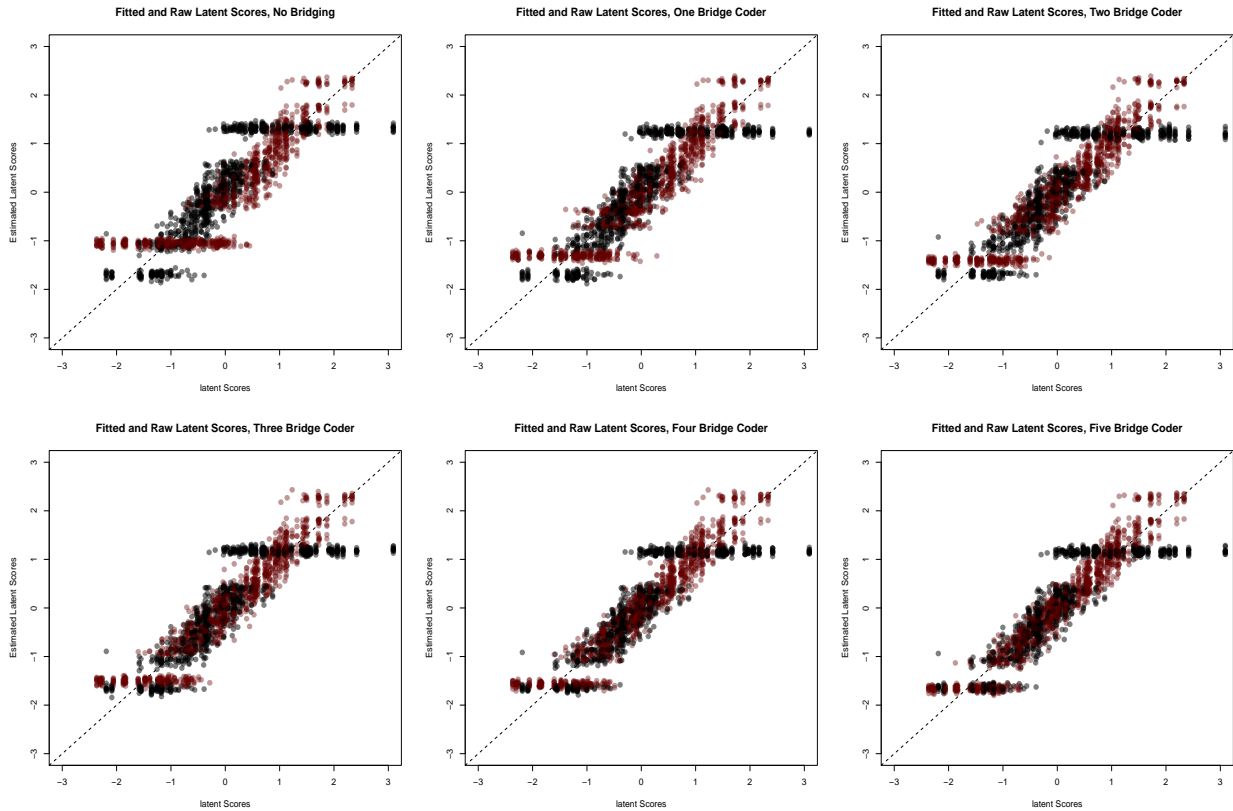


Figure 16: Fitted and real latent score estimates. Bridging is from a random country with low coders' thresholds to a random country with high coders' thresholds. Countries coded by low threshold coders are represented by black dots. Countries coded by high threshold coders are represented by red dots.

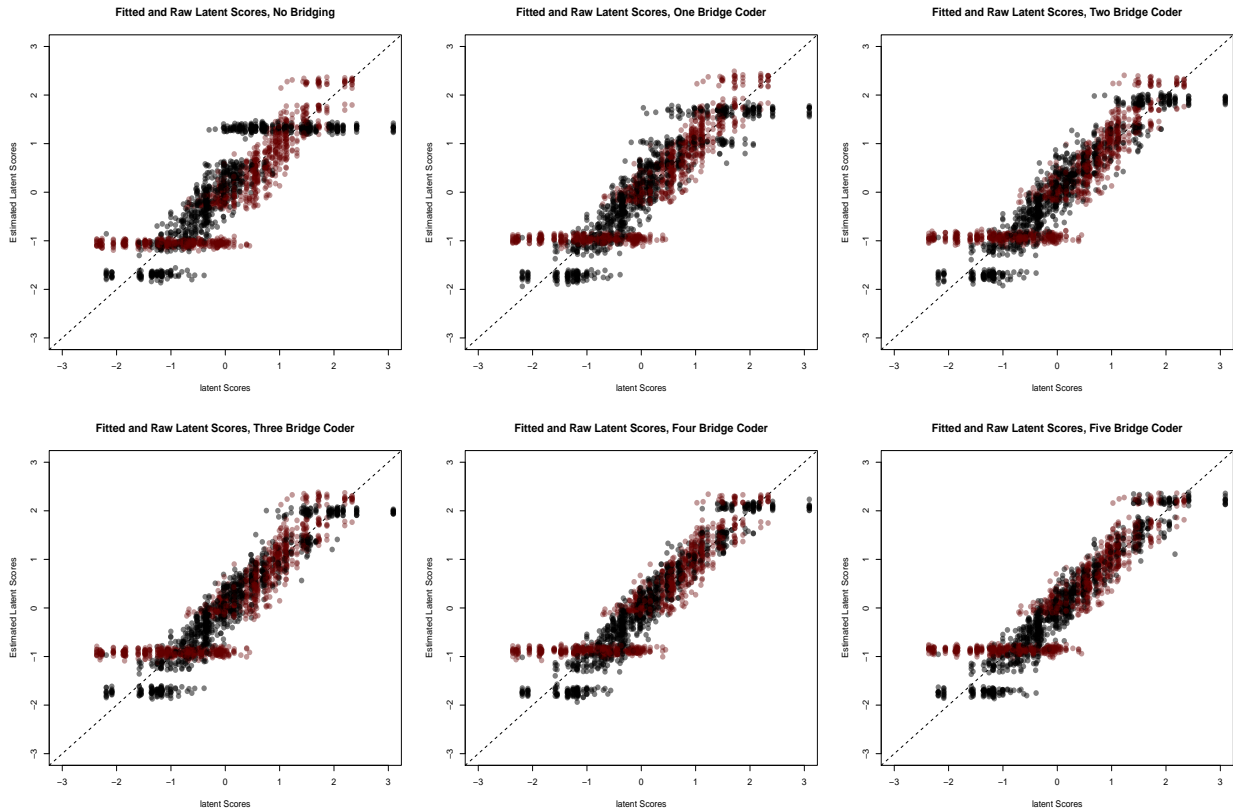


Figure 17: Fitted and real latent score estimates. Bridging is from a random country with high coders' thresholds to a random country with low coders' thresholds. Countries coded by low threshold coders are represented by black dots. Countries coded by high threshold coders are represented by red dots.

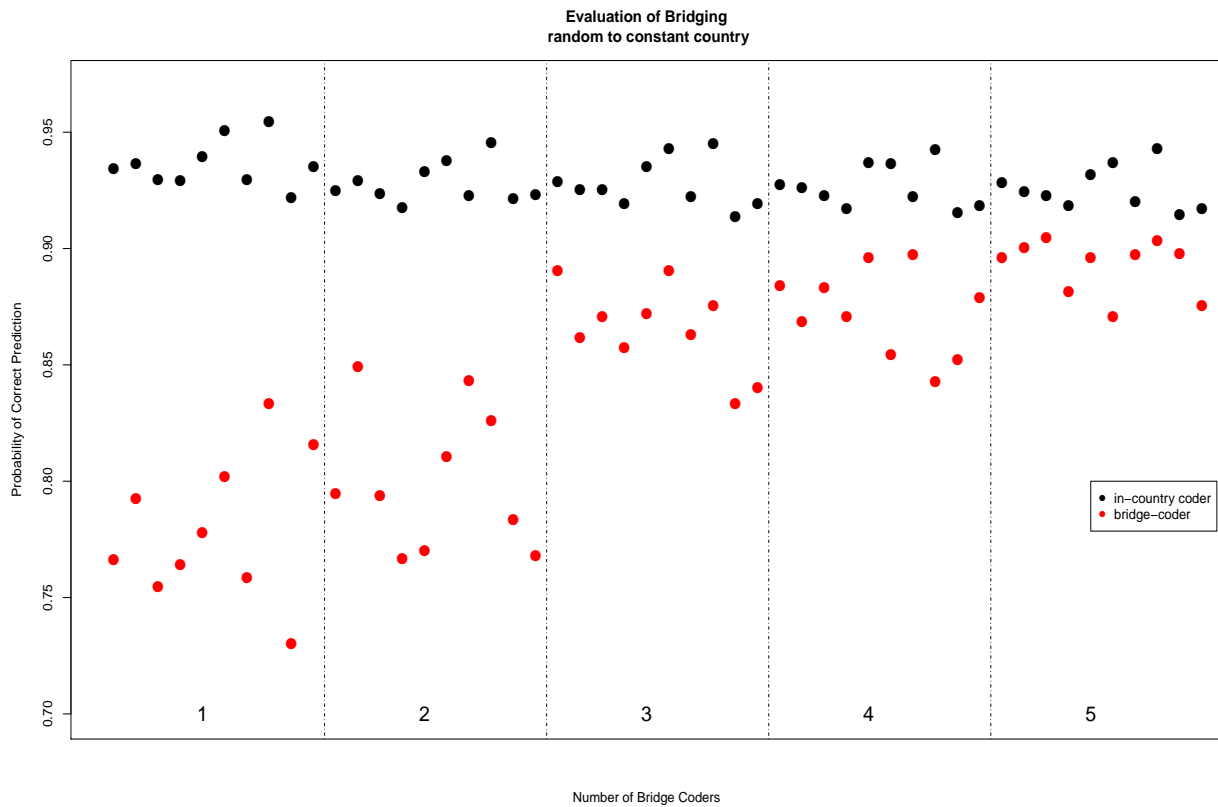


Figure 18: Comparison of posterior predictive probabilities for in-country coders and bridge coders. Bridging is from random country to constant country

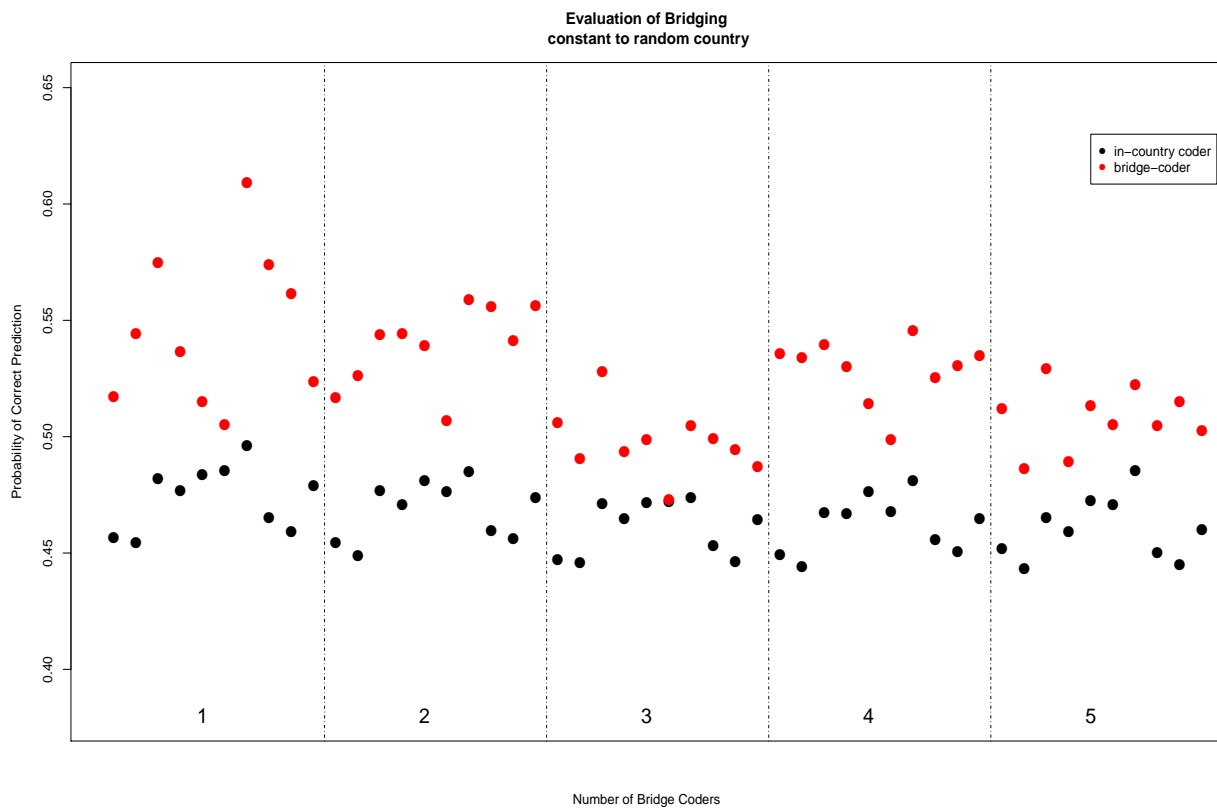


Figure 19: Comparison of posterior predictive probabilities for in-country coders and bridge coders. Bridging is from constant country to random country

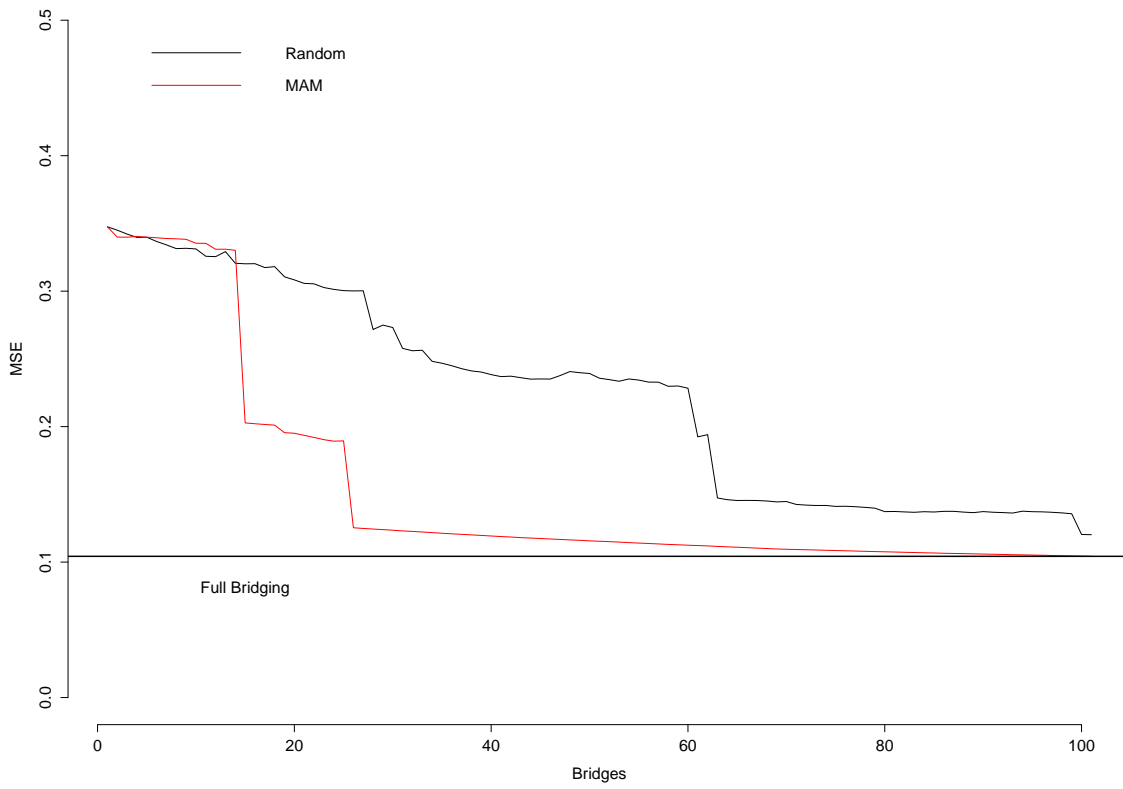


Figure 20: Comparing MAM to random selection.

8 Tables

Table 1: Sums of Mean Square Error

	constant country	random country
c-r bridging	1.13 (0.01)	0.60 (0.009)
r-c bridging	1.60 (0.009)	0.97 (0.01)

Note: Sums of mean square errors for model including full bridging pattern (bottom-right panels of Figures 10 and 11) . c-r bridging implies that coders from constant countries also code random the country. r-c stands for the opposite pattern.

References

- Bafumi, Joseph, Andrew Gelman, David K. Park & Noah Kaplan. 2005. "Practical Issues in Implementing and Understanding Bayesian Ideal Point Estimation." *Political Analysis* 13(2):171–187.
- Bailey, Michael a. 2007. "Comparable Preference Estimates across Time and Institutions for the Court, Congress, and Presidency." *American Journal of Political Science* 51(3):433–448.
- Bakker, Ryan, Catherine de Vries, Erica Edwards, Liesbet Hooghe, Seth Jolly, Gary Marks, Jonathan Polk, Jan Rovny, Marco Steenbergen & Milada Vachudova. forthcoming. "Measuring Party Positions in Europe: The Chapel Hill Expert Survey Trend File, 1999-2010." *Party Politics* .
- Choi, Seung W & Richard J Swartz. 2009. "Comparison of CAT Item Selection Criteria for Polytomous Items." *Applied psychological measurement* 33(6):419–440.
- Clinton, Joshua D. & David E. Lewis. 2007. "Expert Opinion, Agency Characteristics, and Agency Preferences." *Political Analysis* 16(1):3–20.

- Clinton, Joshua D. & John S. Lapinski. 2006. "Measuring Legislative Accomplishment, 1877-1994." *American Journal of Political Science* 50(1):232-249.
URL: <http://doi.wiley.com/10.1111/j.1540-5907.2006.00181.x>
- Fariss, Christopher J. 2014. "Respect for Human Rights has Improved Over Time: Modeling the Changing Standard of Accountability." *American Political Science Review* 108(02):297-318.
- Fish, M. Steven & Matthew Kroenig. 2009. *The Handbook of National Legislatures: A Global Survey*. New York: Cambridge University Press.
- Grimmer, J. 2010. "An Introduction to Bayesian Inference via Variational Approximations." *Political Analysis* 19(1):32-47.
URL: <http://pan.oxfordjournals.org/cgi/doi/10.1093/pan/mpq027>
- Groseclose, Tim, Steven D. Levitt & James Snyder. 1999. "Comparing Interest Group Scores across Time and Chambers: Adjusted ADA Scores for the U.S. Congress." *American Political Science Review* 93(1):33-50.
- Ho, Daniel E. & Kevin M Quinn. 2011. "Sparse Data Item Response Theory Modeling with Interest Group Political Positions."
- Johnson, Valen E & James H Albert. 1999. *Ordinal Data Modeling*. New York: Springer.
- Jordan, MI, Z Ghahramani, TS Jaakkola & LK Saul. 1999. "An introduction to variational methods for graphical models." *Machine learning* 37:183-233.
URL: <http://link.springer.com/article/10.1023/A:1007665907178>
- King, G. & J. Wand. 2006. "Comparing Incomparable Survey Responses: Evaluating and Selecting Anchoring Vignettes." *Political Analysis* 15(1):46-66.
URL: <http://pan.oxfordjournals.org/cgi/doi/10.1093/pan/mpl011>

- Kitschelt, Herbert. 2013. *Dataset of the Democratic Accountability and Linkages Project (DALP)*.
- Klingemann, Hans-Dieter, Andrea Volkens, Judith Bara, Ian Budge & Michael D. McDonald. 2006. *Mapping Policy Preferences II: Estimates for Parties, Electors, and Governments in Eastern Europe, European Union, and OECD 1990-2003*. Oxford: Oxford University Press.
- Konig, T., M. Marbach & M. Osnabrugge. 2013. "Estimating Party Positions across Countries and Time—A Dynamic Latent Variable Model for Manifesto Data." *Political Analysis* 21(4):468–491.
URL: <http://pan.oxfordjournals.org/cgi/doi/10.1093/pan/mpt003>
- Lee, Jerry W., Patricia S. Jones, Yoshimitsu Mineyama & Xinwei Esther Zhang. 2002. "Cultural Differences in Responses to a Likert Scale." *Research in Nursing and Health* 25(4):295–306.
- Linzer, Drew A & Jeffrey K Staton. 2012. "A Measurement Model for Synthesizing Multiple Comparative Indicators: The Case of Judicial Independence."
- Marshall, Monty G., Keith Jagers & Ted Robert Gurr. 2011. *Polity IV Project. Center for Systemic Peace: Polity IV Project*. College Park, MD: Center for International Development and Conflict Management: University of Maryland.
- Martin, Andrew D. & Kevin M. Quinn. 2002. "Dynamic Ideal Point Estimation via Markov Chain Monte Carlo for the U.S. Supreme Court, 1953-1999." *Political Analysis* 10(2):134–153.
- McCarty, Nolan M. & Keith T. Poole. 1995. "Veto Power and Legislation: An Empirical Analysis of Executive and Legislative Bargaining from 1961 to 1986." *Journal of Law, Economics, and Organization* 11(2):282–312.

- Montgomery, J. M. & J. Cutler. 2013. "Computerized Adaptive Testing for Public Opinion Surveys." *Political Analysis* 21(2):172–192.
URL: <http://pan.oxfordjournals.org/cgi/doi/10.1093/pan/mps060>
- Ormerod, J. T. & M. P. Wand. 2010. "Explaining Variational Approximations." *The American Statistician* 64(2):140–153.
URL: <http://www.tandfonline.com/doi/abs/10.1198/tast.2010.09058>
- Pemstein, Daniel, Stephen A Meserve & James Melton. 2010. "Democratic Compromise: A Latent Variable Analysis of Ten Measures of Regime Type." *Political Analysis* 18(4):426–449.
- Poole, Keith T & Howard Rosenthal. 1997. *Congress: A Political-Economic History of Roll Call Voting*. Oxford: Oxford University Press.
- Schnakenberg, Keith & Christopher J Fariss. 2014. "Dynamic Patterns of Human Rights Practices." *Political Science Research and Methods* 2(1):1–31.
- Shor, Boris & Nolan McCarty. 2011. "The Ideological Mapping of American Legislatures." *American Political Science Review* 105(3):530–551.
- Spirling, Arthur & Iain McLean. 2007. "UK OC OK? Interpreting Optimal Classification Scores for the U.K. House of Commons." *Political Analysis* 15(1):161–85.
- Treier, Shawn & Simon Jackman. 2008. "Democracy as a Latent Variable." *American Journal of Political Science* 52(1):201–217.

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 $\{\boldsymbol{\theta}_1, \dots, \boldsymbol{\theta}_M\}$, $q(\boldsymbol{\theta})$ 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, \dots, q_M for minimizing the Kullback-Leiber divergence between q and p are

$$q_i^*(\boldsymbol{\theta}_i) \propto \exp [E_{-\boldsymbol{\theta}_i} \log p(\boldsymbol{\theta}_i|\boldsymbol{\theta}_{-i}, \mathbf{y})]. \quad (7)$$

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).

A.1 A Mean Field Approximation Algorithm for the O-IRT Model

We develop a mean approximation algorithm for the O-IRT model described by equation 3, subject to the following priors:

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

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

$$\begin{aligned} q_{\tilde{y}_{ctr}}^*(\tilde{y}_{ctr}) &\propto \exp[E_{-\tilde{y}_{ctr}} \log p(\tilde{y}_{ctr}|\boldsymbol{\theta}_{-\tilde{y}_{ctr}}, \mathbf{y}_{ctr})] \\ &\sim \mathcal{TN}(E(z_{ct})E(\beta_r), 1, E(\gamma_{r,y_{ctr}-1}), E(\gamma_{r,y_{ctr}})), \\ q_{z_{ct}}^*(z_{ct}) &\propto \exp[E_{-z_{ct}} \log p(z_{ct}|\boldsymbol{\theta}_{-z_{ct}}, \mathbf{y}_{ct})] \\ &\sim \mathcal{N}\left(\frac{\sum_{r \in R_{ct}} E(\beta_r)E(\tilde{y}_{ctr})}{1 + \sum_{r \in R_{ct}} E(\beta_r^2)}, \frac{1}{1 + \sum_{r \in R_{ct}} E(\beta_r^2)}\right), \\ q_{\beta_r}^*(\beta_r) &\propto \exp[E_{-\beta_r} \log p(\beta_r|\boldsymbol{\theta}_{-\beta_r}, \mathbf{y}_r)] \\ &\sim \mathcal{N}\left(\frac{\sum_{c,t \in J_r} E(z_{ct})E(\tilde{y}_{ctr}) + \frac{\mu_\beta}{\sigma_\beta^2}}{\frac{1}{\sigma_\beta^2} + \sum_{c,t \in J_r} E(z_{ct}^2)}, \frac{1}{\frac{1}{\sigma_\beta^2} + \sum_{c,t \in J_r} E(z_{ct}^2)}\right), \\ q_{\gamma_{r,k}}^*(\gamma_{r,k}) &\propto \exp[E_{-\gamma_{r,k}} \log p(\gamma_{r,k}|\boldsymbol{\theta}_{-\gamma_{r,k}}, \mathbf{y}_r)] \\ &\sim U\left(\max\left[-2, \max_{y_{ctr}=k} E(\tilde{y}_{ctr})\right], \min\left[\min_{y_{ctr}=k+1} E(\tilde{y}_{ctr}), 2\right]\right). \end{aligned} \tag{8}$$

The expectations in the approximating distributions are

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

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 8.