Studying Party Systems and Institutions with Expert Surveys

Dan Pemstein

NDSU NORTH DAKOTA STATE UNIVERSITY

V-Dem's Party System Measures

Coming soon: Party cyber-security capacity, hate speech, false information dissemination

Measuring Latent Variables Cross-Nationally

- Coding team
  - Training
  - Focus on inter-coder reliability

Measuring Latent Variables Cross-Nationally

- Coding team
  - Training
  - Focus on inter-coder reliability
- Country experts
  - Often one, or few, per county
  - Limited inter-coder reliability checking

Measuring Latent Variables Cross-Nationally

V-Dem
VARIETIES OF DEMOCRACY

- Coding team
  - Training
  - Focus on inter-coder reliability
- Country experts
  - Often one, or few, per county
  - Limited inter-coder reliability checking
- Some guy you know
  - You
  - Your RA
  - Another researcher

Measuring Latent Variables Cross-Nationally

- Coding team

- Country experts

- Some guy you know

Measuring Latent Variables Cross-Nationally

- Coding team
  - Cross-national consistency
  - Lack of local knowledge
  - Can measure inter-coder reliability
  - Aggregation a minor issue
  - Rarely obtain confidence estimates
- Country experts

- Some guy you know

Measuring Latent Variables Cross-Nationally

- Coding team
  - Cross-national consistency
  - Lack of local knowledge
  - Can measure inter-coder reliability
  - Aggregation a minor issue
  - Rarely obtain confidence estimates
- Country experts
  - Deep case knowledge
  - Lack cross-case comparability
  - Withinin case reliability checks
  - Aggregation challenges
  - Confidence estimates are crude
- Some guy you know

Measuring Latent Variables Cross-Nationally

V-Dem
VARIETIES OF DEMOCRACY

- Coding team
  - Cross-national consistency
  - Lack of local knowledge
  - Can measure inter-coder reliability
  - Aggregation a minor issue
  - Rarely obtain confidence estimates
- Country experts
  - Deep case knowledge
  - Lack cross-case comparability
  - Withinin case reliability checks
  - Aggregation challenges
  - Confidence estimates are crude

- Some guy you know
  - Cross-national consistency
  - Lack of local knowledge
  - No way to document rater reliability
  - No aggregation necessary
  - No systematic estimates of confidence

Measuring Latent Variables Cross-Nationally

Can we leverage local expertise and obtain cross-national comparability?

Measuring Latent Variables Cross-Nationally

Can we leverage local expertise and obtain cross-national comparability?

1. Use local experts and cross-national coders
2. Ask regional experts to code a handful of cases
3. Ask raters to code neighbors/hubs
4. Ask all experts to complete anchoring vignettes
5. Apply item response theory models to
   - Leverage different types of knowledge
   - Systematically aggregate ratings within cases
   - Anchor cases to the same scale
   - Estimate rater reliability
   - Estimate confidence in ratings

Varieties of Democracy Project

- Regime characteristics (mostly *de facto* institutions)
  - 7 "properties" (electoral, liberal, majoritarian, . . .)
  - 329 indicators (152 objective, 177 subjective)
  - 5+ coders/data point for subjective measures
  - Whole world ($\approx$ 200 polities), 1900-present
  - $\approx$ 25 million data points

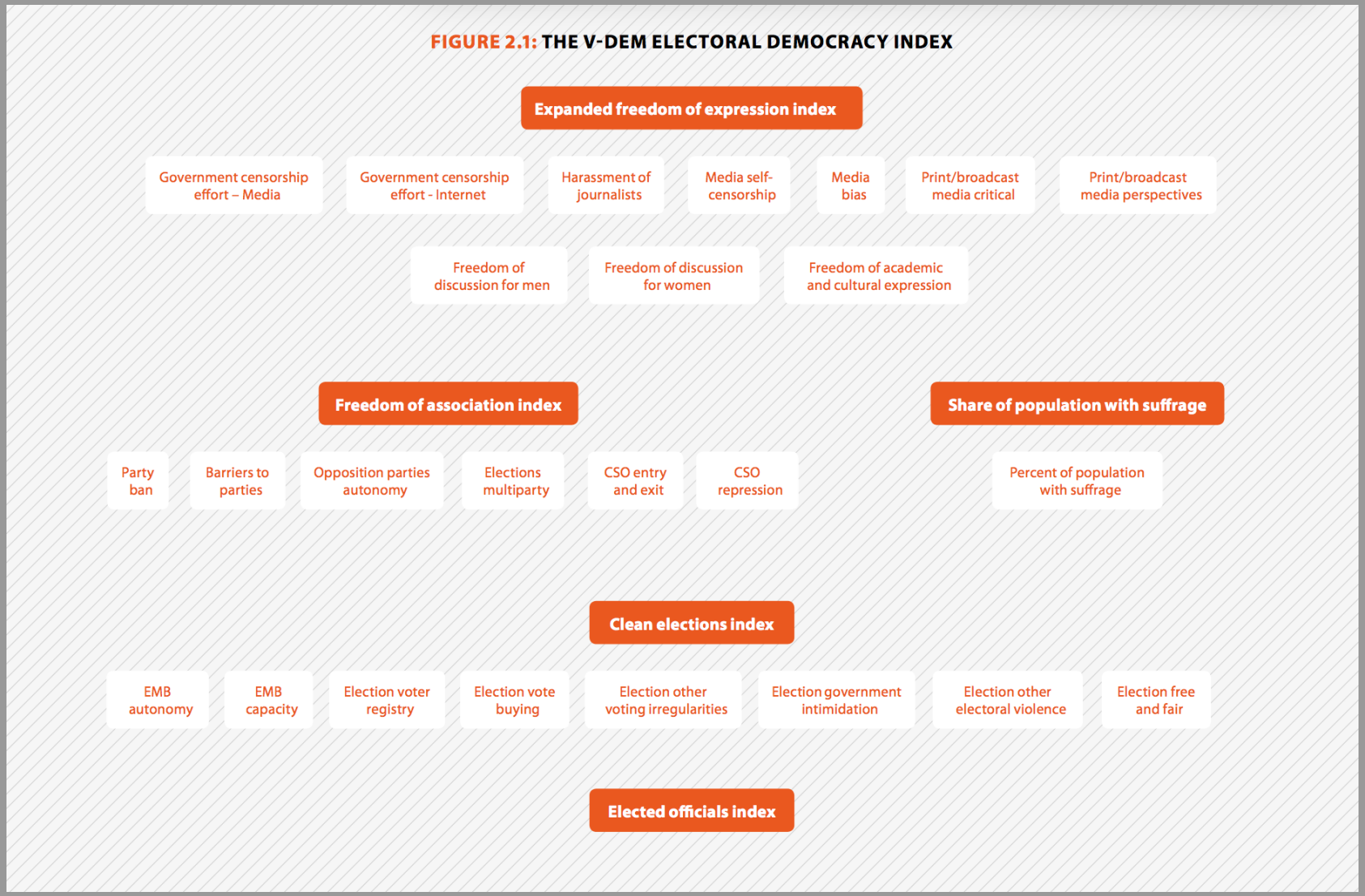
Varieties of Democracy Project

- Regime characteristics (mostly *de facto* institutions)
  - 7 "properties" (electoral, liberal, majoritarian, . . .)
  - 329 indicators (152 objective, 177 subjective)
  - 5+ coders/data point for subjective measures
  - Whole world ($\approx$ 200 polities), 1900-present
  - $\approx$ 25 million data points
- Massive collaborative measurement project
  - 4 PIs, 15 project managers, 3 project coordinators, 32 regional managers, 164 country coordinators, 3500+ country experts, numerous post-docs, RAs, etc
  - Historical extension (1789-1900)

Varieties of Democracy Project

- Regime characteristics (mostly *de facto* institutions)
  - 7 "properties" (electoral, liberal, majoritarian, . . .)
  - 329 indicators (152 objective, 177 subjective)
  - 5+ coders/data point for subjective measures
  - Whole world ($\approx$ 200 polities), 1900-present
  - $\approx$ 25 million data points
- Massive collaborative measurement project
  - 4 PIs, 15 project managers, 3 project coordinators, 32 regional managers, 164 country coordinators, 3500+ country experts, numerous post-docs, RAs, etc
  - Historical extension (1789-1900)
- Uses
  - Aid targeting and assessment
  - Research
    - Democracy and development
    - Comparative institutions
    - Expert knowledge elicitation

V-Dem: A Snapshot



V-Dem: An Example Question

Question: How often do members of the executive (the head of state, the head of government, and cabinet ministers), or their agents, steal, embezzle, or misappropriate public funds or other state resources for personal or family use?

Responses:

0: Constantly. Members of the executive act as though all public resources were their personal or family property.

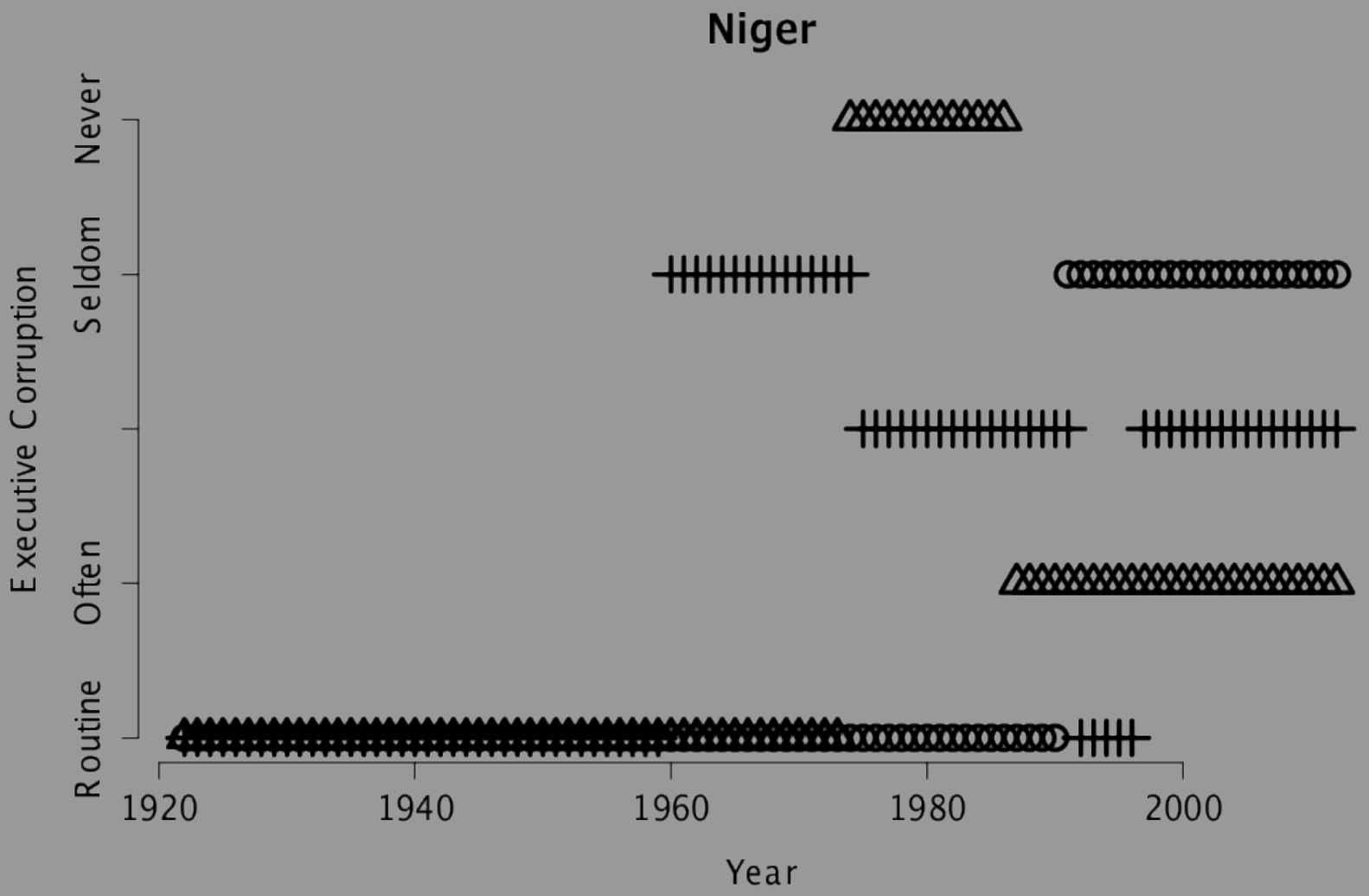
1: Often. Members of the executive are responsible stewards of selected public resources but treat the rest like personal property.

2: About half the time. Members of the executive are about as likely to be responsible stewards of selected public resources as they are to treat them like personal property.

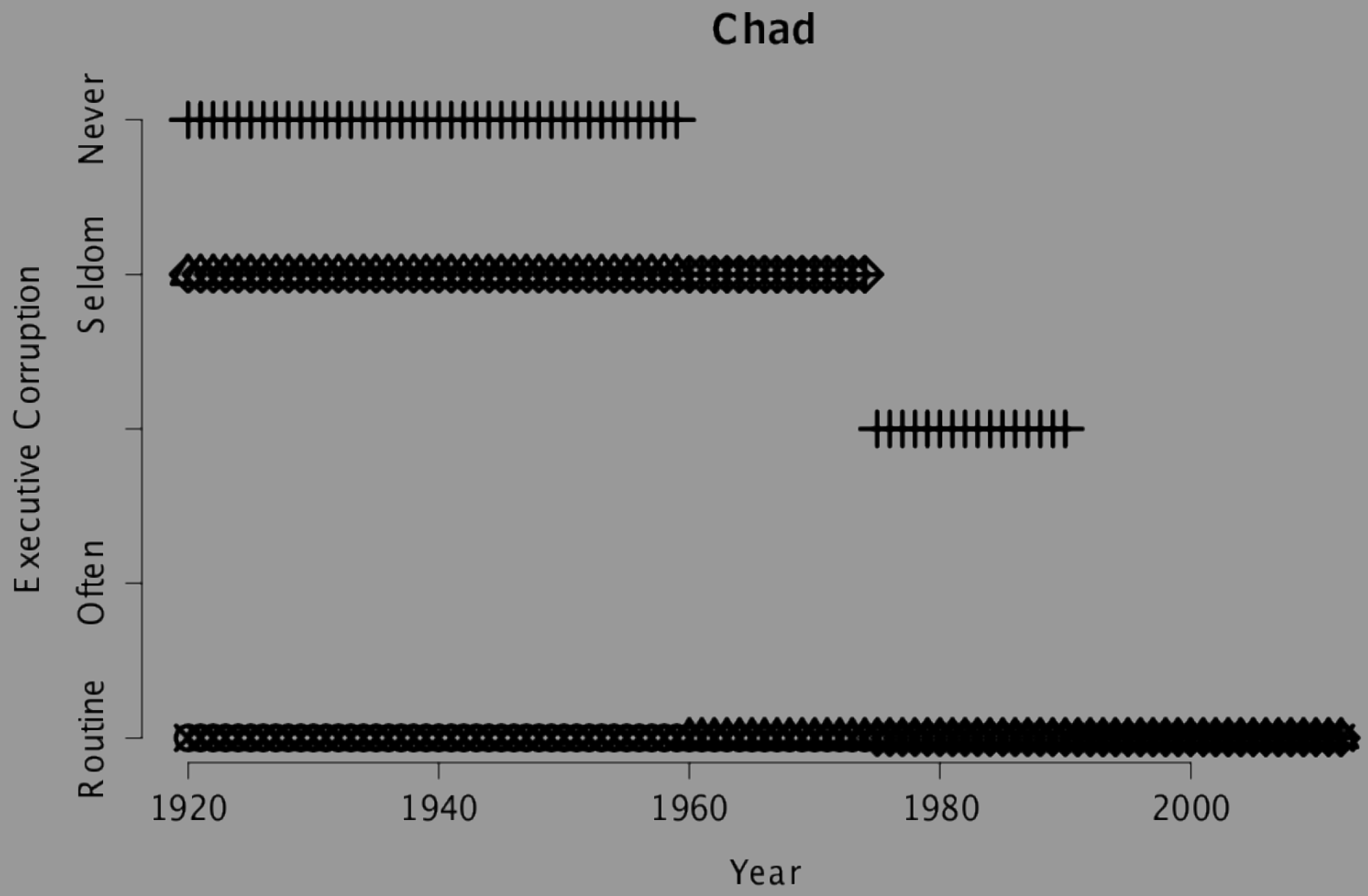
3: Occasionally. Members of the executive are responsible stewards of most public resources but treat selected others like personal property.

4: Never, or hardly ever. Members of the executive are almost always responsible stewards of public resources and keep them separate from personal or family property.

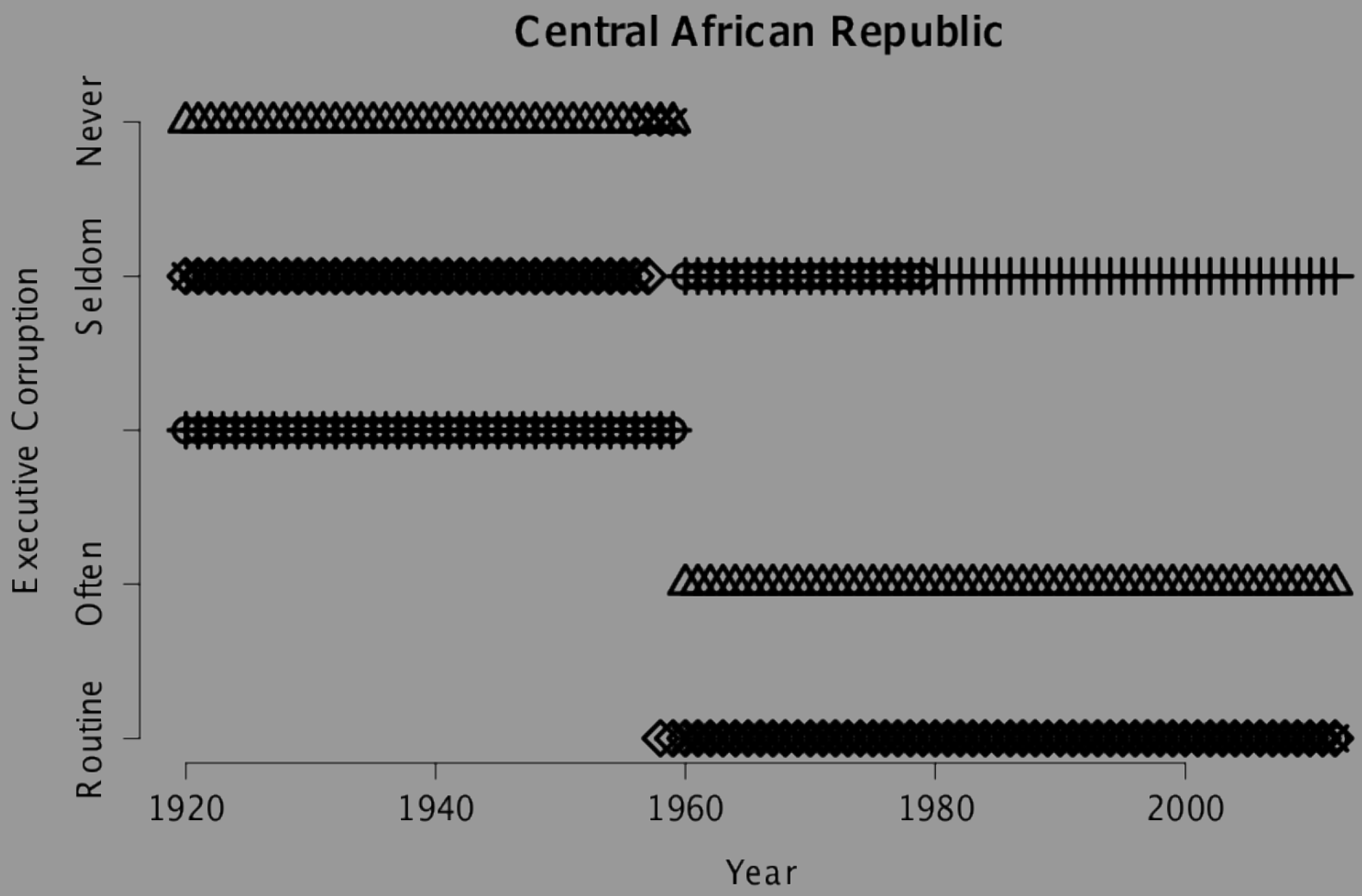
Aggregating Ratings: Raw Data



Aggregating Ratings: Raw Data

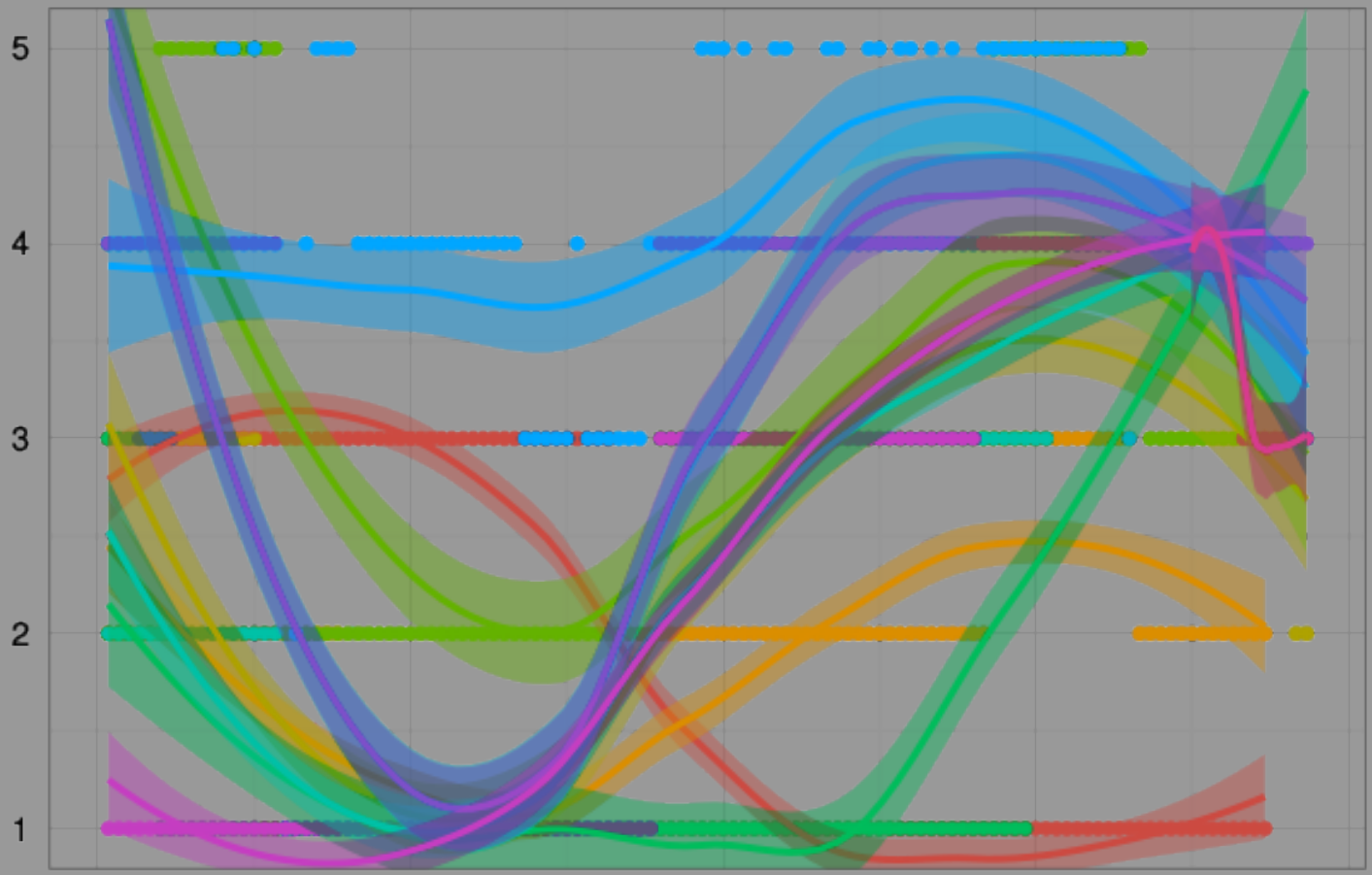


Aggregating Ratings: Raw Data



Problem: DIF & Varying Reliability

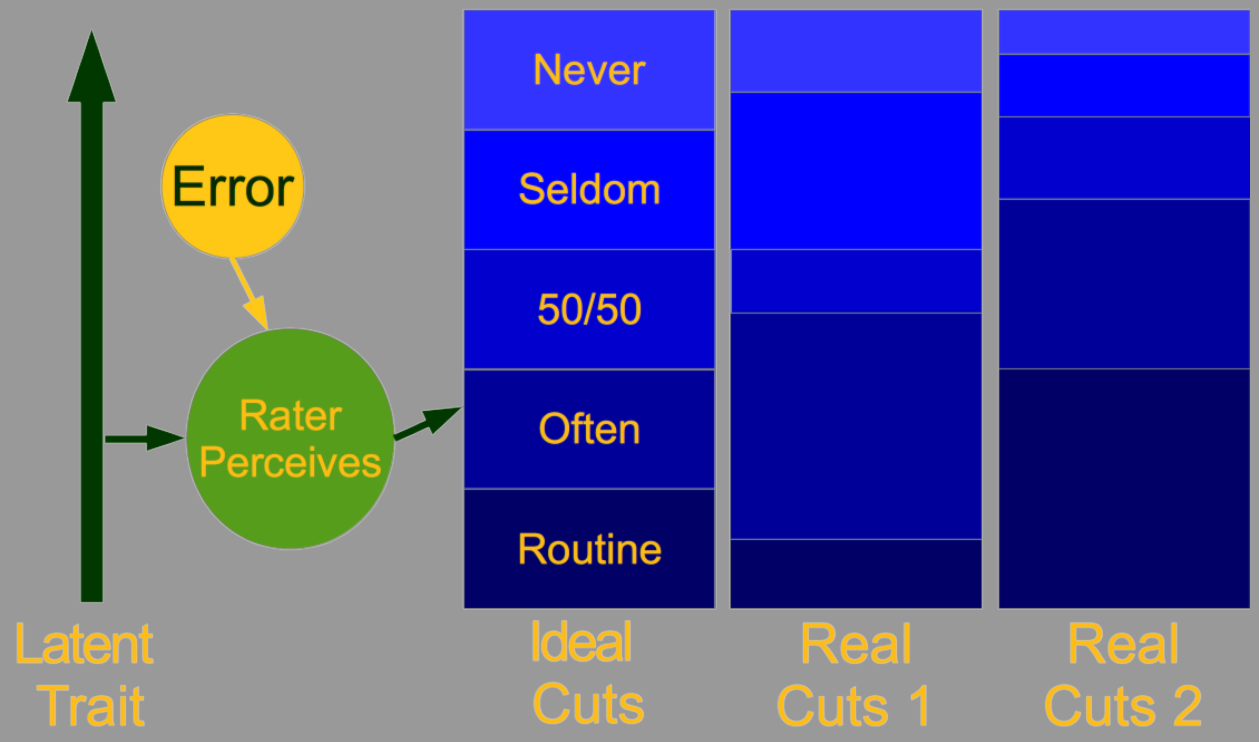
Freedom from political killings in Russia



Aggregating Ratings: DIF & Varying Reliability

Multi-rater ordinal probit / ordinal item response theory model

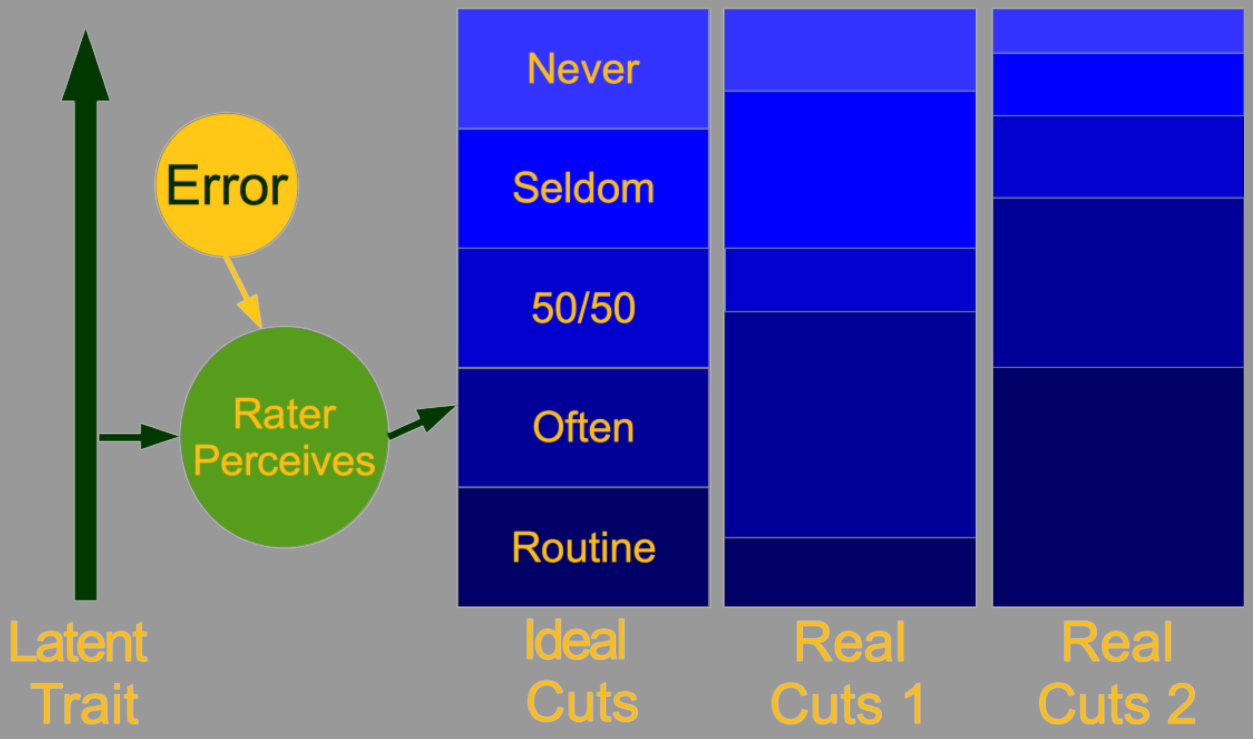
- Rater perceptions equal reality + error
- Raters exhibit arbitrary thresholds
- Given thresholds, raters are correct on average (unbiased)
- Both thresholds and reliability vary across raters



Aggregating Ratings: DIF & Varying Reliability

Bayesian ordinal IRT, assuming $\tilde{y}_{ctr} = z_{ct} + e_{ctr} \sim \mathcal{N}(0, \sigma_r)$

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



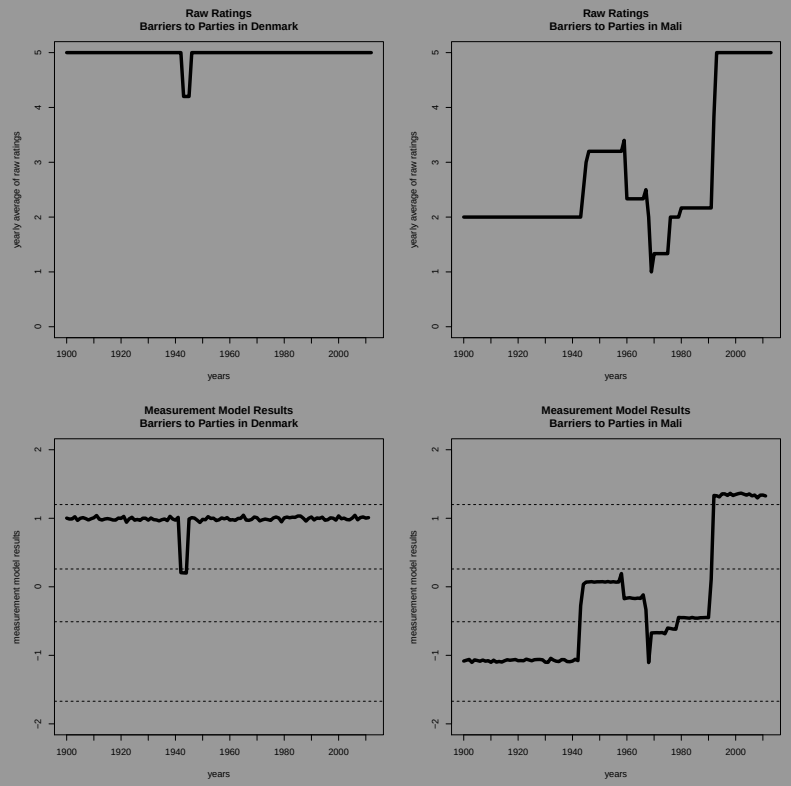
Isn't that a lot of Trouble?

Yes, but:

- Averages much worse with DIF, reliability variance (Marquardt & Pemstein 2018, Lindstädt et al 2018)
- Coverage intervals are also better
- Simple solution (using medians, bootstrapping SEs) is not robust (Marquardt & Pemstein 2019)
- Provides modular framework for dealing with cross-national comparability issues (Bakker & co-authors)

Problem: Cross-National Comparability

- Systematic country-specific DIF
  - Sparse data
  - Like giving a standardized test where kids in every country answer mostly different questions
- Some countries exhibit little variation over coding period



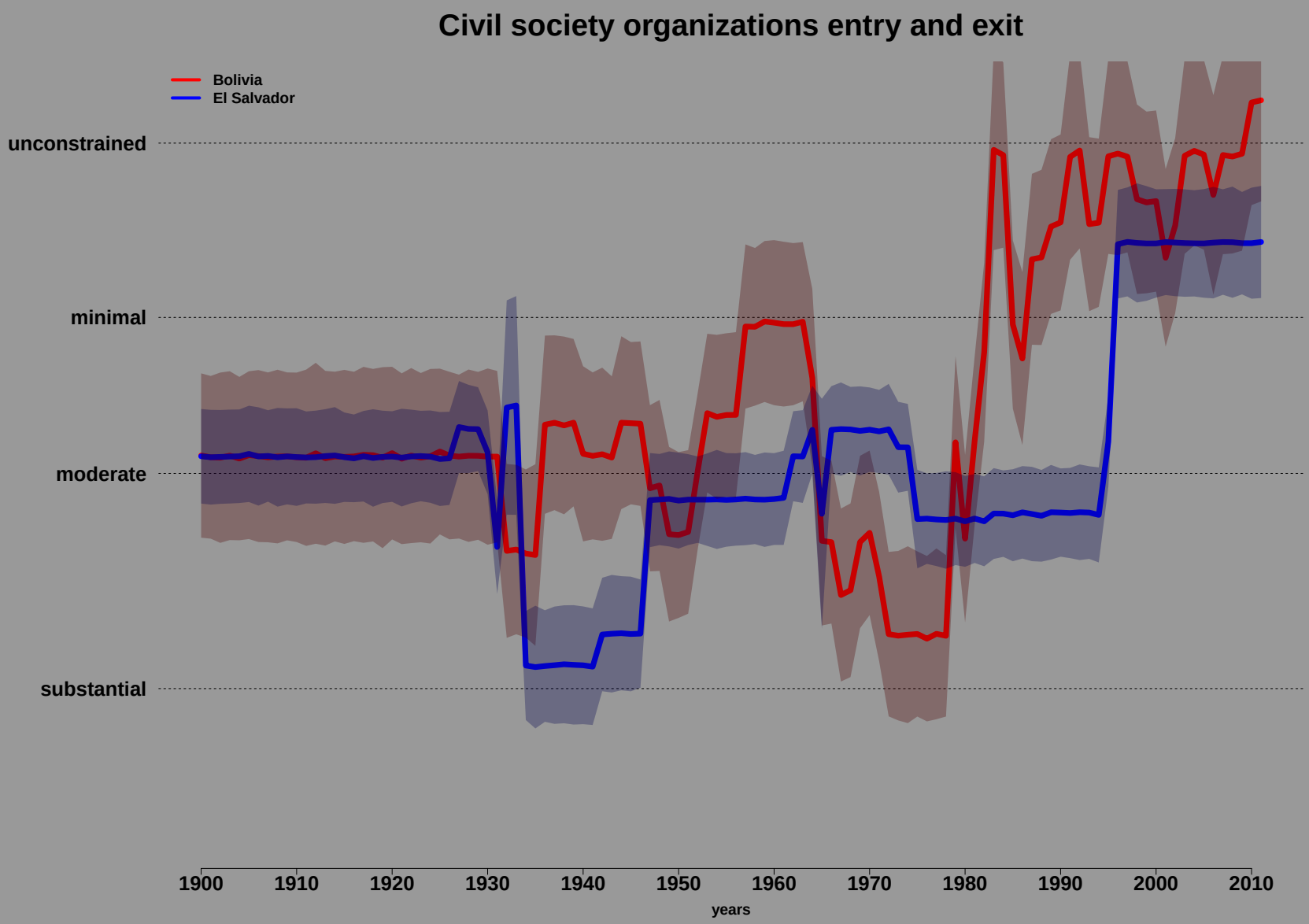
Problem: Cross-National Comparability

- Systematic country-specific DIF
  - Sparse data
  - Like giving a standardized test where kids in every country answer mostly different questions
- Some countries exhibit little variation over coding period
- **Solutions**
  - **Bridge coding**: Coders rate multiple country time series
  - **Lateral coding**: Coders rate multiple countries in a given year
  - **Hierarchical thresholds**: Assume that coders who have a similar coding profile have similar thresholds
  - **Empirical priors**: Ignore DIF when we lack sufficient information to correct for it
  - **Collapse data**: Regime codes amplify lateral coding
  - **Vignettes**

Problem: Cross-National Comparability

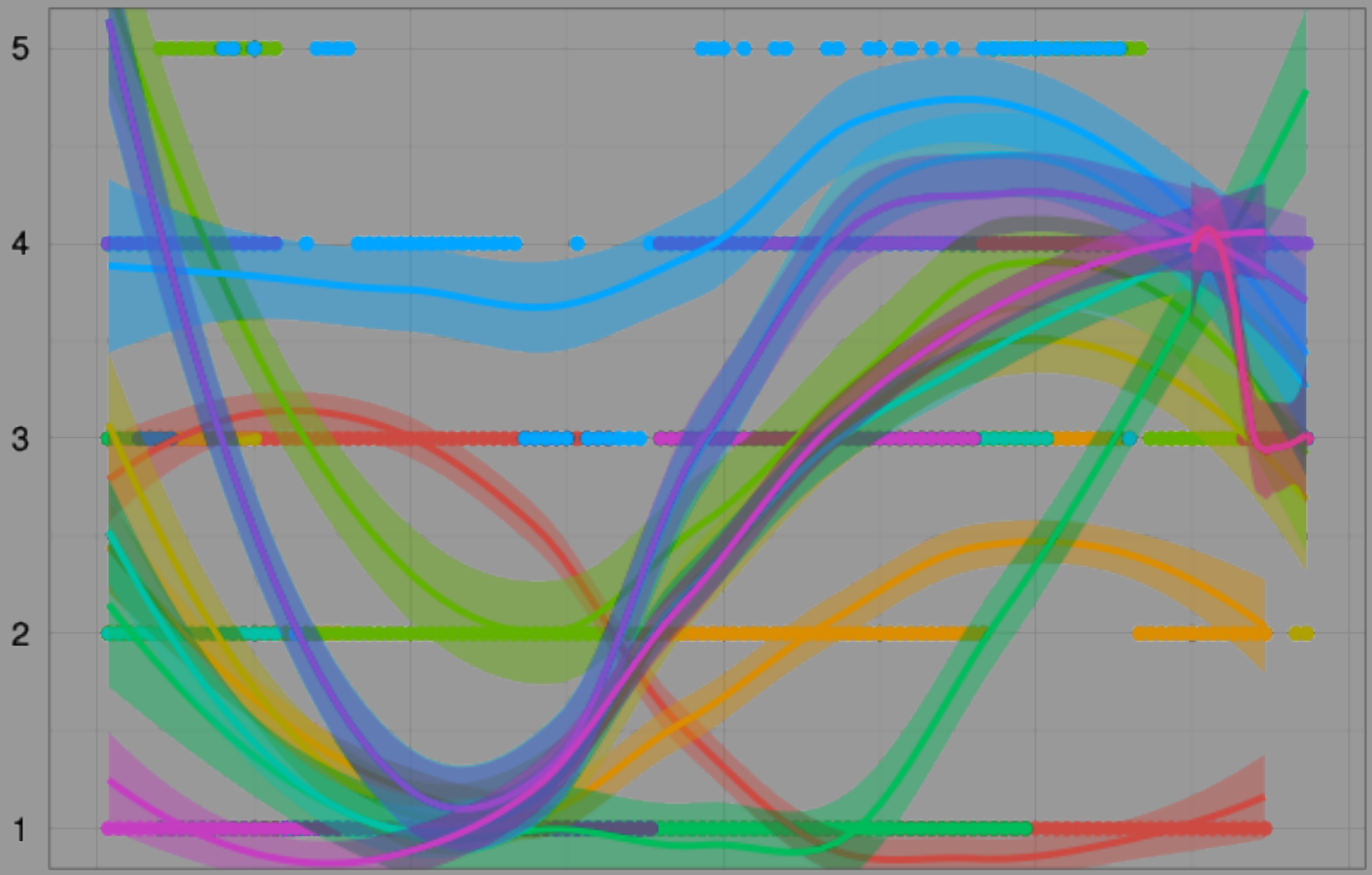
- Systematic country-specific DIF
  - Sparse data
  - Like giving a standardized test where kids in every country answer mostly different questions
- Some countries exhibit little variation over coding period
- **Solutions**
  - **Bridge coding**: Coders rate multiple country time series
  - **Lateral coding**: Coders rate multiple countries in a given year
  - **Hierarchical thresholds**: Assume that coders who have a similar coding profile have similar thresholds
  - **Empirical priors**: Ignore DIF when we lack sufficient information to correct for it
  - **Collapse data**: Regime codes amplify lateral coding
  - **Vignettes**
- Future direction: Expert pairwise comparisons

Latent Trait Estimates



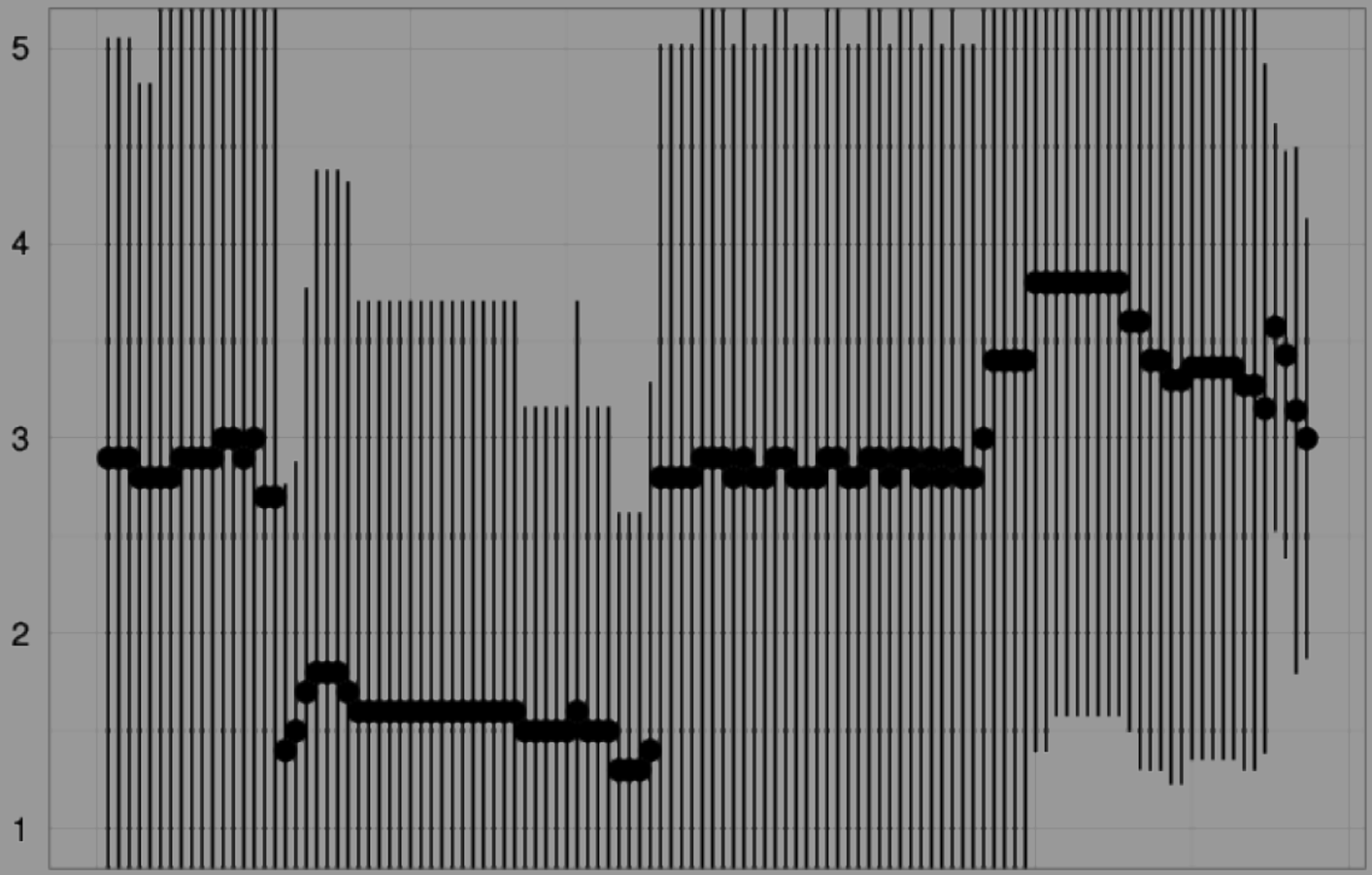
Advantages of Latent Trait Estimates

Freedom from political killings in Russia



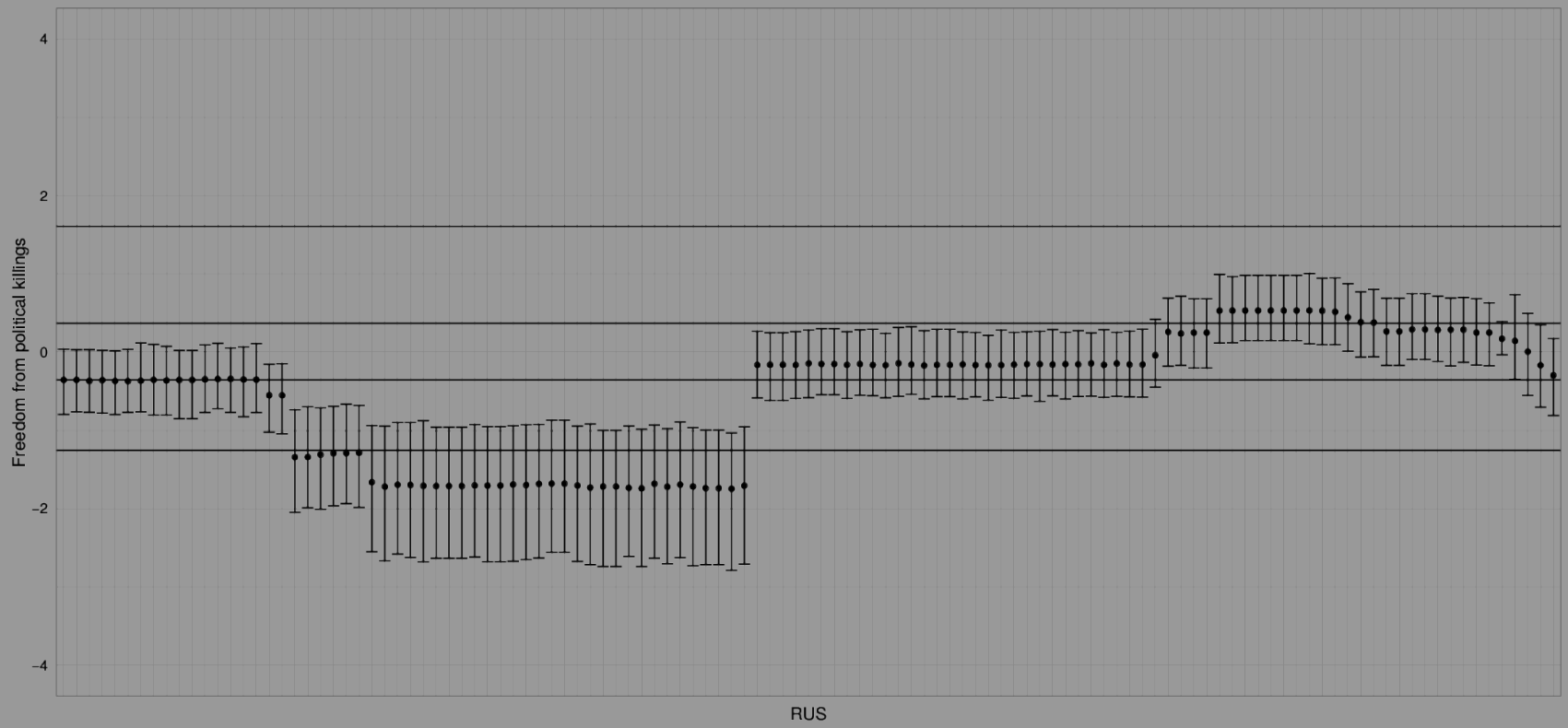
Advantages of Latent Trait Estimates

Freedom from political killings in Russia

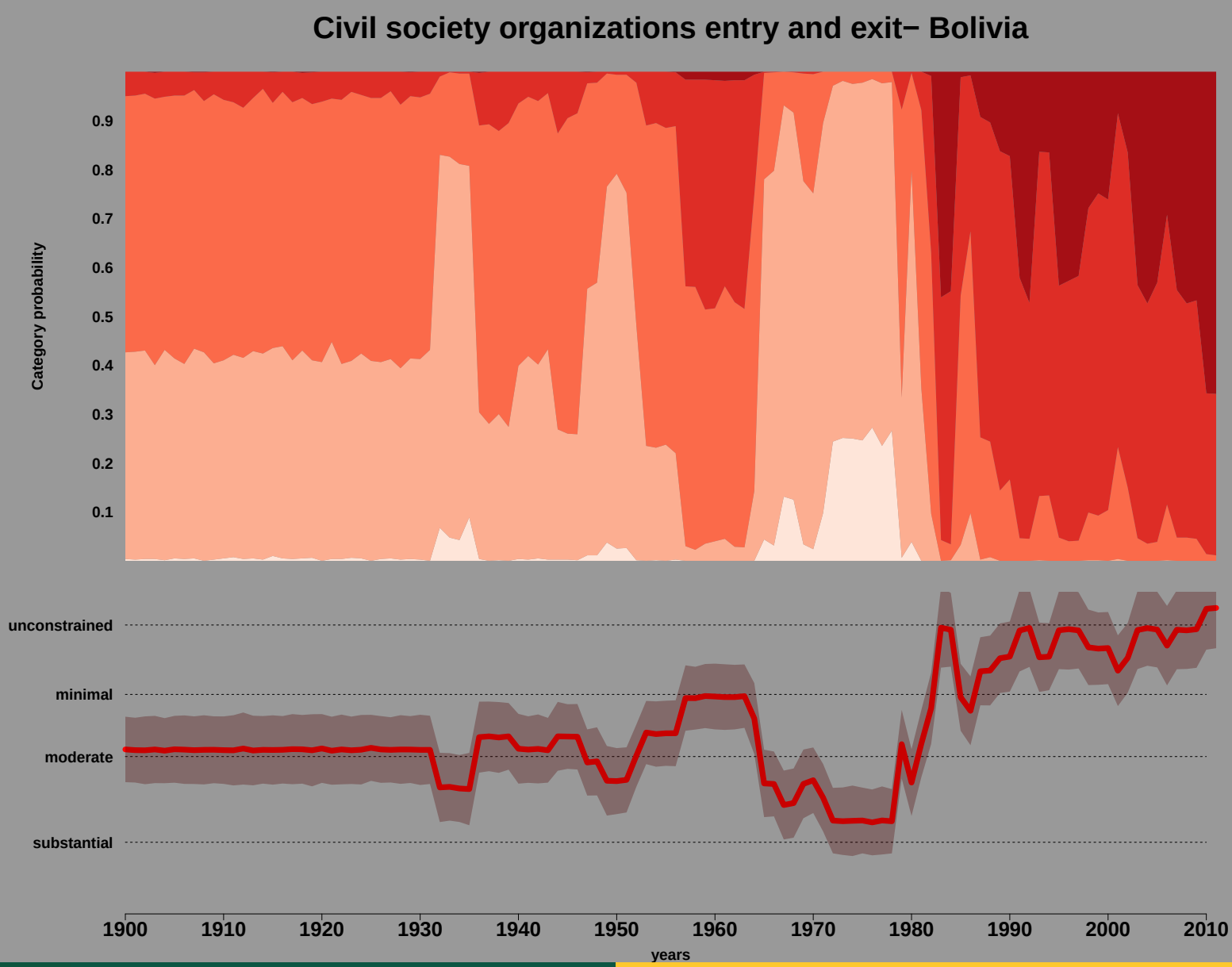


Advantages of Latent Trait Estimates

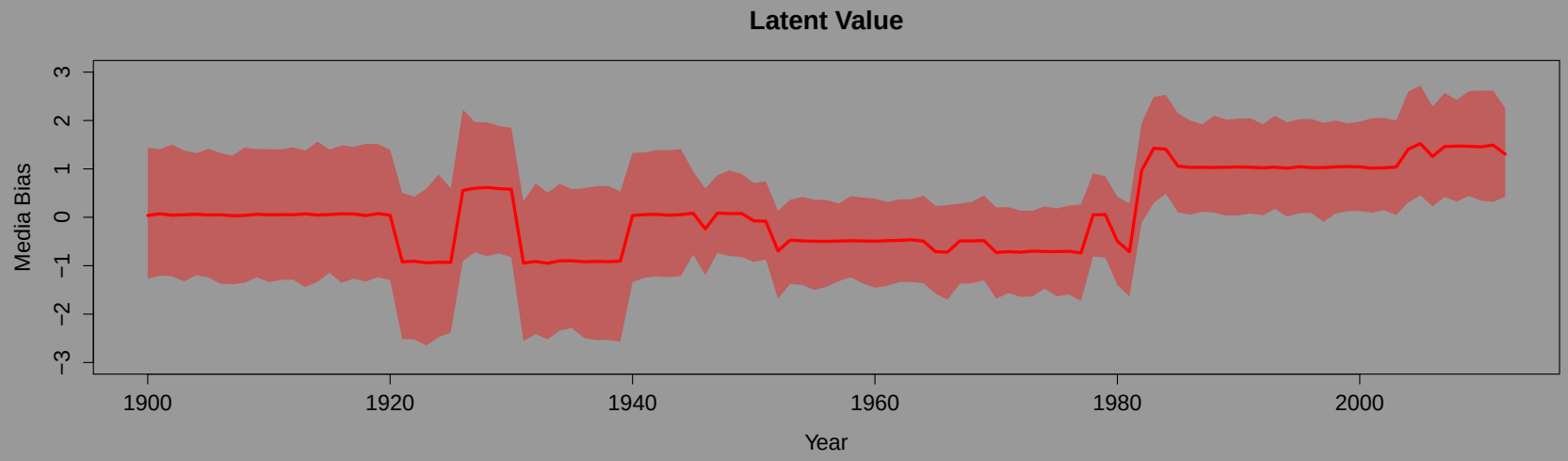
Freedom from political killings in Russia

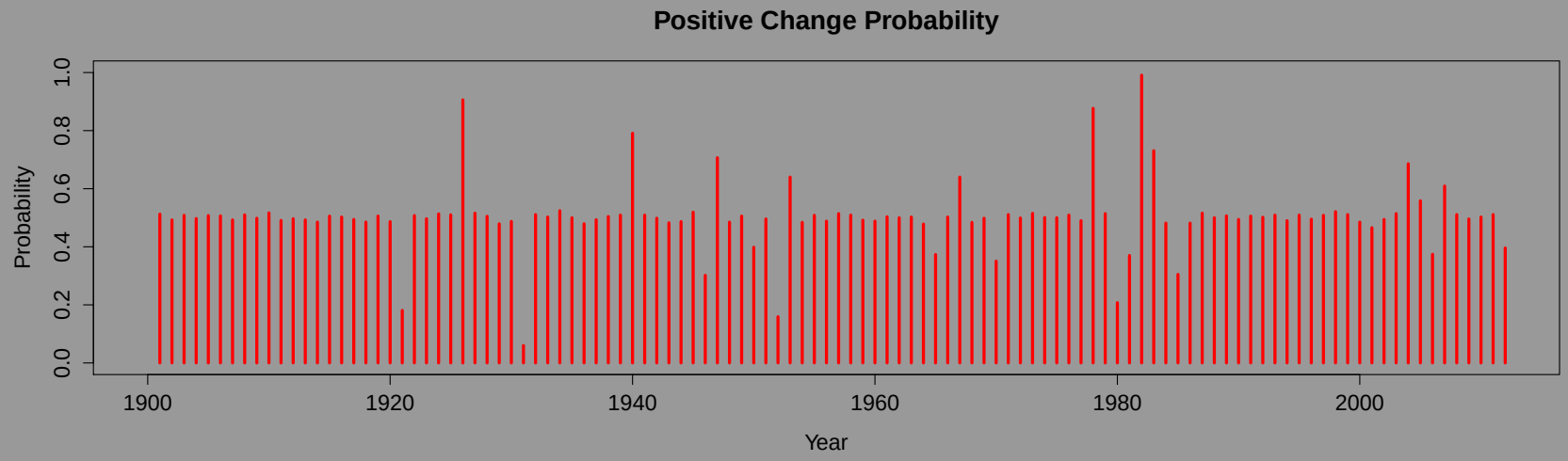


Posterior Ordinal Probabilities

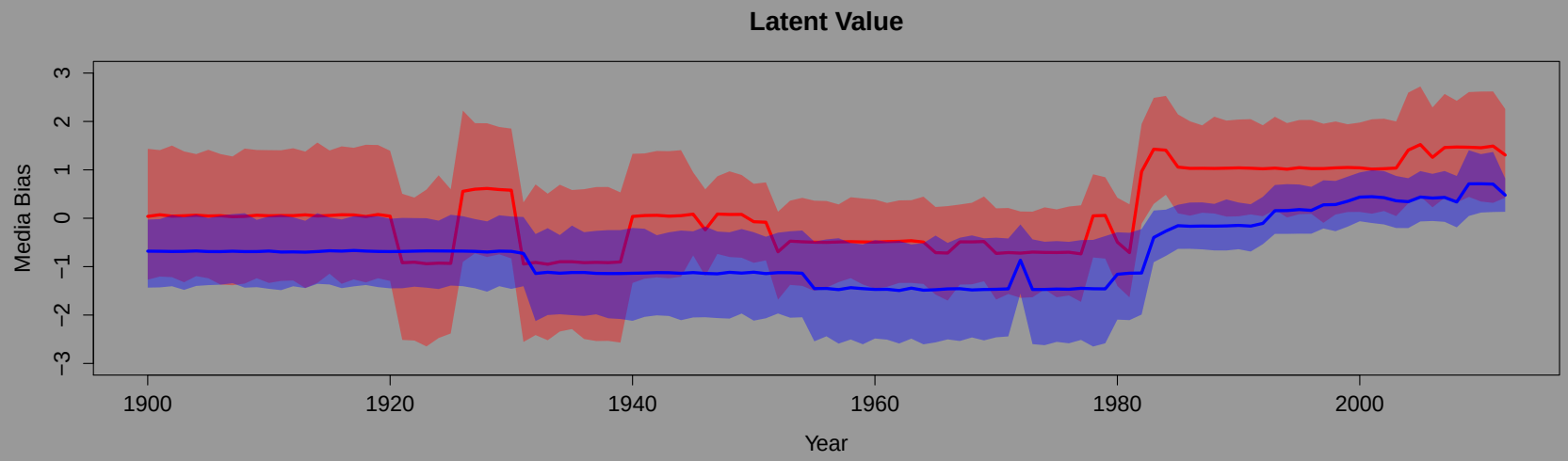


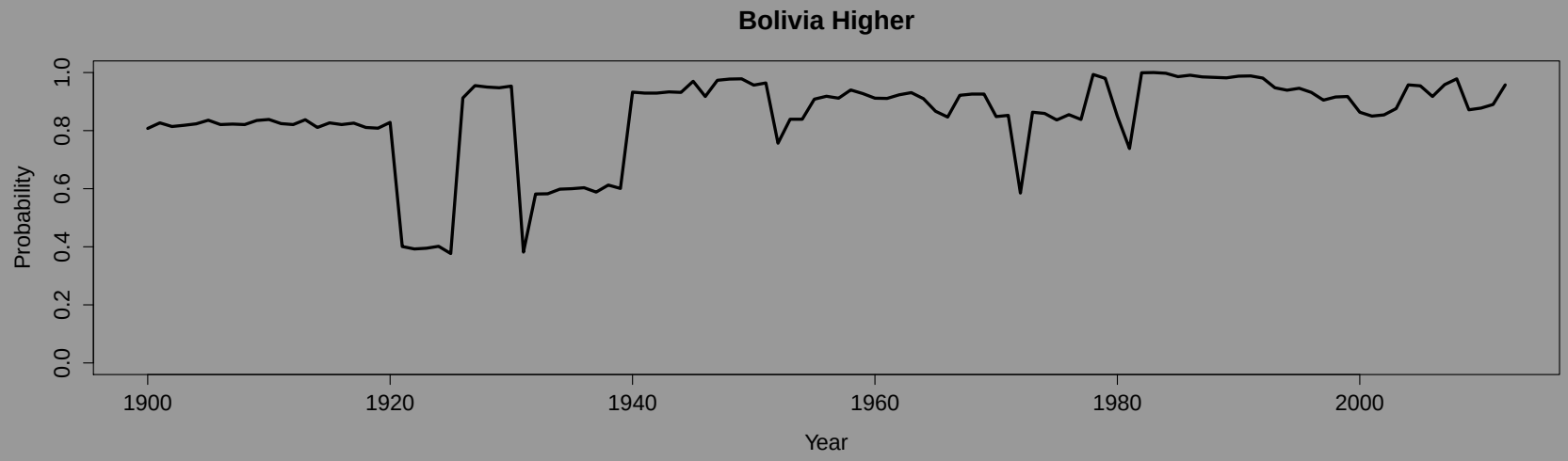
Change Probabilities





Probabilistic Comparisons





Thanks!