Online appendix: "IRT models for expert-coded panel data"

A Freedom from political killings question

Question: Is there freedom from political killings?
Clarification: Political killings are killings by the state or its agents without due process of law for the purpose of eliminating political opponents. These killings are the result of deliberate use of lethal force by the police, security forces, prison officials, or other agents of the state (including paramilitary groups).
Responses:

1: Not respected by public authorities. Political killings are practiced systematically and they are typically incited and approved by top leaders of government.

2: Weakly respected by public authorities. Political killings are practiced frequently and top leaders of government are not actively working to prevent them.

3: Somewhat respected by public authorities. Political killings are practiced occasionally but they are typically not incited and approved by top leaders of government.

4: Mostly respected by public authorities. Political killings are practiced in a few isolated cases but they are not incited or approved by top leaders of government.

5: Fully respected by public authorities. Political killings are non-existent.

Figure A.1: V–Dem Question 10.5, Freedom from Political Killings.

B Data on the characteristics of expert coders

Freedom from political killings is a variable with great variation in expert characteristics. Among the 1,171 unique experts who coded these data there are 164 unique countries-of-birth, 158 unique countries-of-residence, and 128 countries-of-education. Sixty-two percent of the experts hold a PhD, 27 percent an MA, three percent a professional degree (e.g. MD, JD), seven percent a BA or equivalent, and less than one percent just a secondary level

of education or post-secondary vocational training. Sixty-one percent of experts work at a university, 13 percent at an NGO, seven percent are self-employed, six percent are students, three percent work in the private sector, four percent work for a government organ, and 2 percent work for a state-owned enterprise. Twenty-seven percent of experts are female, and the mean age in 2014 was 45. Given this wide variation in backgrounds, there is strong reason to expect that experts would vary in their perceptions of the latent concept.

In terms of variation in expert reliability, experts vary along a variety of factors that may proxy their average expertise. First, there is variation among experts in terms of the number of countries and country-years they code. On average, experts code approximately two unique countries, with a range from one to 29 countries. The average expert codes 78 ($sd = 65$) country-years. Given that experts may become less reliable as they code countries with which they are less familiar, and may experience fatigue the more country-years they code, this variation should yield variance in expert reliability.

Experts also evince variation in the degree to which they vary their codings: the average standard deviation in coding is 0.71 ($sd = 0.54$). While there are many valid reasons why an expert may not vary her coding (e.g. an expert could have only coded countries that did not vary greatly in their scores, such as Switzerland), in many other cases coding variation clearly measures the degree to which an expert was attentive to changes in her country and thus her reliability.

C Bayesian A–M model

Aldrich-McKelvey (A–M) scaling provides an alternative method for converting ordinal data to a latent scale. A–M performs a linear scaling of ordinal data, and accounts for DIF using expert-specific intercept and slope parameters; in Bayesian implementations it can also account for variation in expert reliability in the form of stochastic error variance with an expert-specific variance parameter. We follow Hare et al. (2015) in developing a Bayesian

A–M model (BAM) based on the likelihood in equation C.1.

$$\begin{aligned} y_{ctr} &\sim \mathcal{N}(\mu_{ctr}, \tau_r) \\ \mu_{ctr} &= \alpha_r + \beta_r z_{ct} \end{aligned} \tag{C.1}$$

Here α_r, β_r and τ_r are expert-specific intercept, slope and variance parameters. Thus, BAM replaces the k expert-specific difficulty parameters in the most general threshold-based IRT framework with linear intercept and slope parameters. The expert-specific variance parameter is analogous to the discrimination parameter in an IRT model; both measure rater-specific error variance.

Note that Hare et al. (2015) estimate both the observation and coder precision parameters. But since most country-years in our data have only a handful of coders, this approach is not tenable here.[23] Instead, we estimate τ_r as follows:

$$\begin{aligned} \tau_r^{-1} &\sim \Gamma(\upsilon, \omega) \\ \upsilon &\sim \Gamma(1, 1) \\ \omega &\sim \Gamma(1, 1) \end{aligned} \tag{C.2}$$

For the same reason of data sparsity, we eschew the standard practice of assigning model parameters vague uniform priors. Instead, we assign α_r a $\mathcal{N}(0, 5)$ prior and β_r a Log-Normal prior, $ln(\beta_r) \sim \mathcal{N}(0, ln[2])$.

While more flexible than the intercept-only IRT approach, this model is less general

[23]While parameters appeared to converge across chains in testing runs, according to the standard Gelman-Rubin diagnostic, these runs exhibited large numbers of divergent transitions, a potentially strong indicator of lack of convergence for models fit with Stan. While we fit only a handful of models with this general prior specification, they recovered true values at rates almost identical to the other A–M specifications that we present here.

than models that incorporate threshold DIF: it assumes that DIF only occurs through linear transformations on the latent scale, not non-linearly, at individual thresholds. Fully ordinal IRT models can capture classes of DIF that are assumed away by BAM, and therefore rely on less restrictive assumptions about DIF's functional form. Thus, such models are more robust in principle. On the other hand, BAM's simpler parameterization might provide advantages when dealing with sparse data, since it demands less information than ordinal IRT.

Finally, because we are interested in comparing our hierarchical IRT specifications to currently-used approaches, we focus on a non-hierarchical BAM implementation. To ensure that this distinction does not drive differences between IRT and A–M performance, we also fit a handful of hierarchical A–M models (HAM), as a robustness check.

Our HAM specification closely follows the BAM model. We adopt the same likelihood function, but alter the priors such that

$$\begin{aligned} \beta_r &\sim \mathcal{N}(\beta_{c_r}, 0.11) \\ \beta_{c_r} &\sim \mathcal{N}(\beta_\mu, 0.11) \\ ln(\beta_\mu) &\sim \mathcal{N}(1.5, 2), \end{aligned} \tag{C.3}$$

$$\begin{aligned} \alpha_r &\sim \mathcal{N}(\alpha_{c_r}, 0.13) \\ \alpha_{c_r} &\sim \mathcal{N}(\alpha_\mu, 0.13) \\ \alpha_\mu &\sim \mathcal{N}(3.1, 4), \end{aligned} \tag{C.4}$$

and

$$\tau_r^{-1} \sim \Gamma(1, 1). \tag{C.5}$$

The prior specifications for the α and β parameters follow hierarchical specifications analogous to those for the threshold parameters in the IRT models. Prior means and variances

Bridging	DIF Type	DIF Level	Rel. Var.	MSE	ρ	τ	95% HPD	Div. Trans.
V-Dem	None		Fixed	0.21	0.89	0.71	0.89	0
V-Dem	Intercept	High	High	0.61	0.67	0.49	0.67	10
V-Dem	Threshold	High	High	0.44	0.76	0.56	0.82	0
V-Dem	Truncated	High	High	0.46	0.75	0.56	0.8	0
High	Intercept	High	High	0.31	0.84	0.67	0.69	217
High	Threshold	High	High	0.16	0.92	0.76	0.87	12
High	Truncated	High	High	0.20	0.90	0.73	0.81	6
None	Intercept	High	High	0.68	0.63	0.47	0.64	0
None	Threshold	High	High	0.39	0.78	0.59	0.87	0
None	Truncated	High	High	0.53	0.71	0.52	0.74	635

Table C.1: Hierarchical A–M Performance

are based on actual simulated values in threshold-based simulations with linear threshold steps. In other words, we based these priors on true simulated values, under the A–M linearity assumption. We simplify the prior on τ in order to reduce estimation issues. This prior is less flexible than that in the other A–M model, but places substantial mass over the true τ values in the simulated data. In sum, this hierarchical specification is consistent with the actual simulation process, potentially providing substantial advantages to the HAM model.

Table C.1 presents simulation performance statistics for a subset of simulated datasets to which we fit HAM models.[24] In general, the hierarchical specification does little to improve model fits over vanilla A–M results. While the HAM substantially improves fits for V-Dem bridged data with no DIF and fixed reliability, it generally produces similar performance to the vanilla A–M models. Indeed, the only other substantive improvement in MSE was for no-bridging datasets with high threshold DIF and high reliability variance, although this HAM still under-performed IRT approaches in these data. Nonetheless, adding hierarchical parameters appears to improve HPD interval coverage across most specifications, although not to the extent that HAM models tend to outperform IRT models on this dimension.

[24]To provide a reasonable robustness check, while conserving computational resources, we focus primarily on high DIF/high reliability variance datasets, across different bridging and DIF specifications.

We experienced some computational difficulties fitting the HAM model to some of our simulated datasets. In particular, four of the models produced divergent transitions after burnin. Results from these models may be misleading. Notably, the Gelman-Rubin diagnostic was inconsistent with convergence for the truncated dataset with no bridging, which produced 635 divergent transitions. Even with hierarchical parameters, A–M models are prone to convergence issues when applied to these data.

D STAN code

D.1 Model without DIF or reliability parameters

```
  data {
    int<lower=2> K;//categories
    int<lower=0> J; // Coders
    int<lower=0> N; // N
    int<lower=0> C; // countries
    int<lower=-1,upper=K> wdata[N,J];// data
    int<lower=1,upper=C> cdata[J]; // j country indices
}

  parameters {
    vector[N] Z;
    ordered[K-1] gamma; // world-level cutpoints
  }

  model {
    vector[K] p;
    real left;
    real right;

    for(i in 1:N) {
      Z[i] ~ normal(0, 1);
    }


    for (j in 1:J) {
      for (i in 1:N) if (wdata[i,j] != -1) {
        left <- 0;
```

```
        for (k in 1:(K-1)) {
          right <- left;
          left <- Phi_approx(gamma[k] - Z[i]);
          p[k] <- left - right;
        }
        p[K] <- 1.0 - left;
        wdata[i,j] ~ categorical(p);
      }
    }
  }
```

D.2 Model without DIF and with reliability parameters

```
  data {
    int<lower=2> K;//categories
    int<lower=0> J; // Coders
    int<lower=0> N; // N
    int<lower=0> C; // countries
    int<lower=-1,upper=K> wdata[N,J];// data
    int<lower=1,upper=C> cdata[J]; // j country indices
}

  parameters {
    vector[N] Z;
    ordered[K-1] gamma; // world-level cutpoints
    real<lower=0> beta[J]; //reliability
  }

  model {
    vector[K] p;
    real left;
    real right;

    for(i in 1:N) {
      Z[i] ~ normal(0, 1);
    }

for (j in 1:J) {
      beta[j] ~ normal(1,1)T[0,];
      for (i in 1:N) if (wdata[i,j] != -1) {
        left <- 0;
        for (k in 1:(K-1)) {
          right <- left;
          left <- Phi_approx(gamma[k] - beta[j]*Z[i]);
          p[k] <- left - right;
```

```
      }
      p[K] <- 1.0 - left;
      wdata[i,j] ~ categorical(p);
    }
  }
}
```

D.3 Model with intercept DIF and reliability parameters

```
data {
  int<lower=2> K;//categories
  int<lower=0> J; // Coders
  int<lower=0> N; // N
  int<lower=0> C; // countries
  int<lower=-1,upper=K> wdata[N,J];// data
  int<lower=1,upper=C> cdata[J]; // j country indices
}

parameters {
  vector[N] Z;
  ordered[K-1] gamma; // world-level cutpoints
  vector[C] epsilon_c; // country-level agreement
  real epsilon[J]; //agreement
  real<lower=0> beta[J]; //agreement
}

model {
  vector[K] p;
  real left;
  real right;

  for(i in 1:N) {
    Z[i] ~ normal(0, 1);
  }

for (c in 1:C) {
    epsilon_c[c] ~ normal(0, .5); // row-access of gamma_c
  }

  for (j in 1:J) {
    epsilon[j] ~ normal(epsilon_c[cdata[j]], .5);  // note row-access
    beta[j] ~ normal(1,1)T[0,];
    for (i in 1:N) if (wdata[i,j] != -1) {
      left <- 0;
      for (k in 1:(K-1)) {
```

```
        right <- left;
        left <- Phi_approx(gamma[k] - epsilon[j] - beta[j]*Z[i]);
        p[k] <- left - right;
      }
      p[K] <- 1.0 - left;
      wdata[i,j] ~ categorical(p);
    }
  }
}
```

D.4 Model with threshold DIF and reliability parameters

```
data {
  int<lower=2> K;//categories
  int<lower=0> J; // Coders
  int<lower=0> N; // N
  int<lower=0> C; // countries
  int<lower=-1,upper=K> wdata[N,J];// data
  int<lower=1,upper=C> cdata[J]; // j country indices
}

parameters {
  vector[N] Z;
  ordered[K-1] gamma[J];
  vector[K-1] gamma_mu; // world-level cutpoints
  matrix[C, (K-1)] gamma_c; // country-level cuts, rows are countries
  real<lower=0> beta[J]; //reliability score
}

model {
  vector[K] p;
  real left;
  real right;

  for(i in 1:N) {
    Z[i] ~ normal(0, 1);
  }
  gamma_mu ~ uniform(-2, 2);

  for (c in 1:C) {
    gamma_c[c] ~ normal(gamma_mu, .25); // row-access of gamma_c
  }

  for (j in 1:J) {
    gamma[j] ~ normal(gamma_c[cdata[j]], .25);  // note row-access
```

```
      beta[j] ~ normal(1,1)T[0,];

      for (i in 1:N) if (wdata[i,j] != -1) {
        left <- 0;
        for (k in 1:(K-1)) {
          right <- left;
          left <- Phi_approx(gamma[j,k] - Z[i]*beta[j]);
          p[k] <- left - right;
        }
        p[K] <- 1.0 - left;
        wdata[i,j] ~ categorical(p);
      }
    }
  }
```

D.5 BAM model

```
  data {
    int<lower=0> J; // Coders
    int<lower=0> N; // N
    int<lower=-1,upper=5> wdata[N,J];// data
}

  parameters {
    vector[N] Z;
    real<lower=0> tau[J]; //reliability
    real<lower=0> beta[J]; //reliability
    vector[J] alpha; //reliability
    real<lower=0> a; //reliability
    real<lower=0> b; //reliability
  }

  model {

    a ~ gamma(1,1);
    b ~ gamma(1,1);
    for(i in 1:N) {
      Z[i] ~ normal(0, 1);
    }

for (j in 1:J) {
      beta[j] ~ lognormal(0,log(2));
      alpha[j] ~ normal(0,5);
      tau[j] ~ gamma(a,b);
```

```
    for (i in 1:N) if (wdata[i,j] != -1) {
      wdata[i,j] ~ normal(alpha[j] + beta[j]*Z[i], 1/tau[j]);
    }
  }
}
```

D.6 HAM model

```
data {
  int<lower=0> J; // Coders
  int<lower=0> N; // N
  int<lower=0> C; // countries
  int<lower=-1,upper=5> wdata[N,J];// data
  int<lower=1,upper=C> cdata[J]; // j country indices
}

parameters {
  vector[N] Z;
  real<lower=0> tau[J]; //reliability
  real<lower=0> beta[J];
  real<lower=0> beta_c[C];
  real<lower=0> beta_mu;
  vector[J] alpha;
  vector[C] alpha_c;
  real alpha_mu;
  //real<lower=0> a;
  //real<lower=0> b;
}

model {

  //a ~ gamma(1,1);
  //b ~ gamma(1,1);
  for(i in 1:N) {
    Z[i] ~ normal(0, 1);
  }

  alpha_mu ~ normal(3.1, 4);
  beta_mu ~ lognormal(log(1.5), log(2));

  for (c in 1:C) {
    alpha_c[c] ~ normal(alpha_mu, 0.11);
    beta_c[c] ~ normal(beta_mu, 0.13);
  }
```

```
  for (j in 1:J) {
    beta[j] ~ normal(beta_c[cdata[j]], 0.11);
    alpha[j] ~ normal(alpha_c[cdata[j]], 0.13);
    //tau[j] ~ gamma(a,b);
    tau[j] ~ gamma(1,1);

    for (i in 1:N) if (wdata[i,j] != -1) {
      wdata[i,j] ~ normal(alpha[j] + beta[j]*Z[i], 1/tau[j]);
    }
  }
}
```

E Additional illustrative cases of different IRT models

Table E.1: Correlation between median estimates across different models, actual V–Dem data

Model	Normalized mean	No DIF		Intercept DIF		Hierarchical DIF		A–M algorithm
	Mean	Fixed	Varying	Fixed	Varying	Fixed	Varying	
Normalized mean	1	.97	.81	.75	.72	.90	.82	.65
No DIF, fixed reliability	.99	1	.82	.75	.72	.89	.82	.65
No DIF, varying reliability	.95	.95	1	.68	.73	.78	.88	.59
Intercept DIF, fixed reliability	.91	.92	.87	1	.83	.82	.72	.73
Intercept DIF, varying reliability	.89	.90	.90	.96	1	.77	.76	.70
Hierarchical DIF, fixed reliability	.97	.99	.94	.96	.93	1	.82	.69
Hierarchical DIF, varying reliability	.95	.95	.98	.90	.92	.96	1	.63
A–M algorithm	.83	.82	.78	.89	.88	.86	.82	1

Pearson correlation coefficients below the diagonal, Kendall correlation coefficients above the diagonal.

Table E.2: Correlation between median estimates across different models, V–Dem observations with more than five coders

Model	Normalized mean	No DIF		Intercept DIF		Hierarchical DIF		A–M algorithm
	Mean	Fixed	Varying	Fixed	Varying	Fixed	Varying	
Normalized mean	1	.97	.82	.79	.75	.90	.82	.68
No DIF, fixed reliability	.99	1	.83	.79	.76	.90	.83	.67
No DIF, varying reliability	.95	.95	1	.71	.76	.79	.89	.61
Intercept DIF, fixed reliability	.93	.94	.90	1	.84	.85	.75	.74
Intercept DIF, varying reliability	.92	.92	.92	.96	1	.80	.78	.72
Hierarchical DIF, fixed reliability	.98	.99	.94	.97	.94	1	.83	.71
Hierarchical DIF, varying reliability	.95	.96	.98	.92	.94	.96	1	.65
A–M algorithm	.85	.84	.80	.90	.89	.87	.84	1

Pearson correlation coefficients below the diagonal, Kendall correlation coefficients above the diagonal.

Table E.3: Correlation between median estimates across different models, V–Dem observations with fewer than five coders

Model	Normalized mean	No DIF		Intercept DIF		Hierarchical DIF		A–M algorithm
	Mean	Fixed	Varying	Fixed	Varying	Fixed	Varying	
Normalized mean	1	.95	.81	.71	.67	.89	.82	.59
No DIF, fixed reliability	.99	1	.82	.70	.66	.87	.82	.57
No DIF, varying reliability	.94	.95	1	.65	.67	.76	.82	.51
Intercept DIF, fixed reliability	.87	.88	.84	1	.83	.79	.72	.70
Intercept DIF, varying reliability	.85	.86	.86	.96	1	.73	.75	.65
Hierarchical DIF, fixed reliability	.97	.98	.92	.94	.91	1	.82	.65
Hierarchical DIF, varying reliability	.95	.95	.94	.90	.92	.96	1	.57
A–M algorithm	.75	.74	.68	.87	.82	.81	.75	1

Pearson correlation coefficients below the diagonal, Kendall correlation coefficients above the diagonal.

Figure E.1: Different models of freedom from political killings in Germany

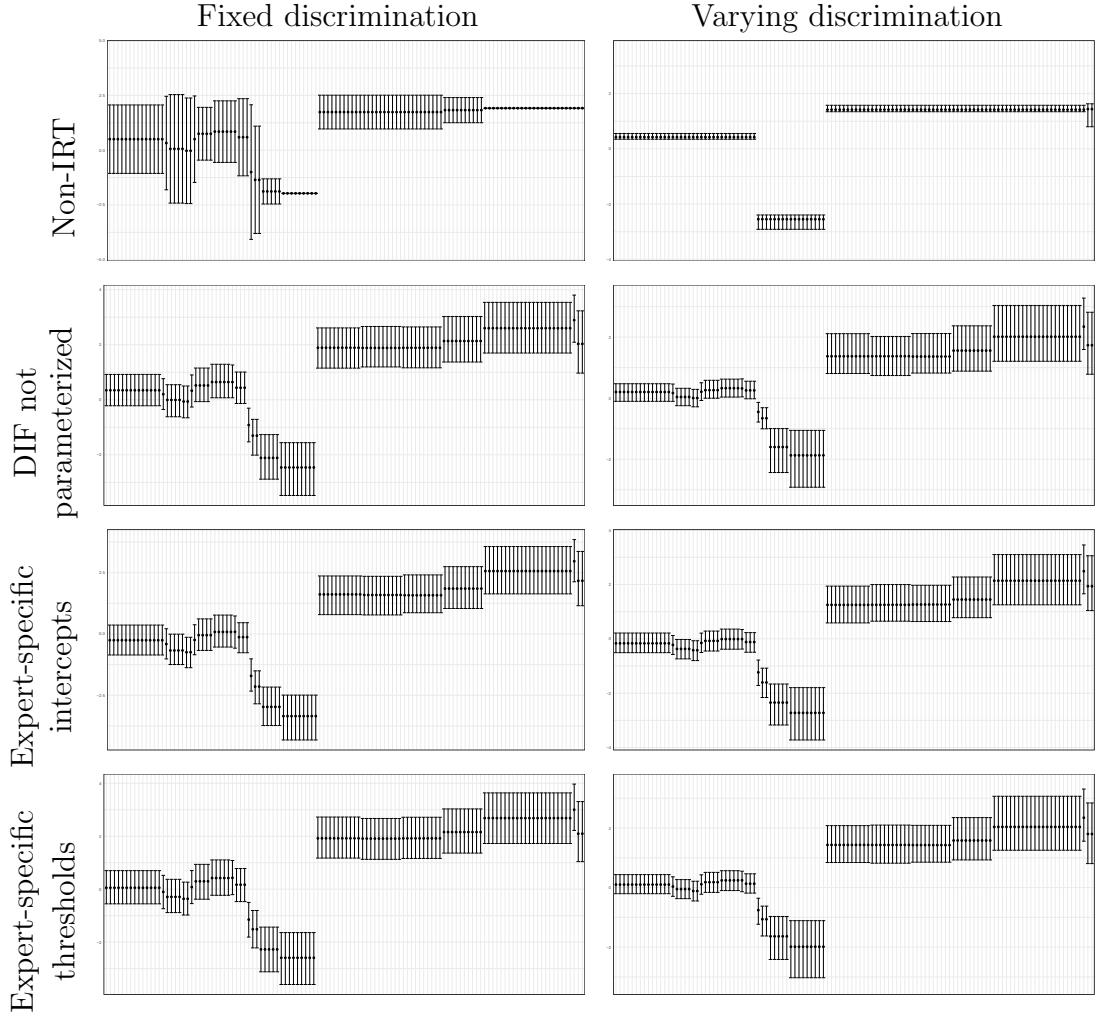


Figure E.2: Different models of freedom from political killings in Canada

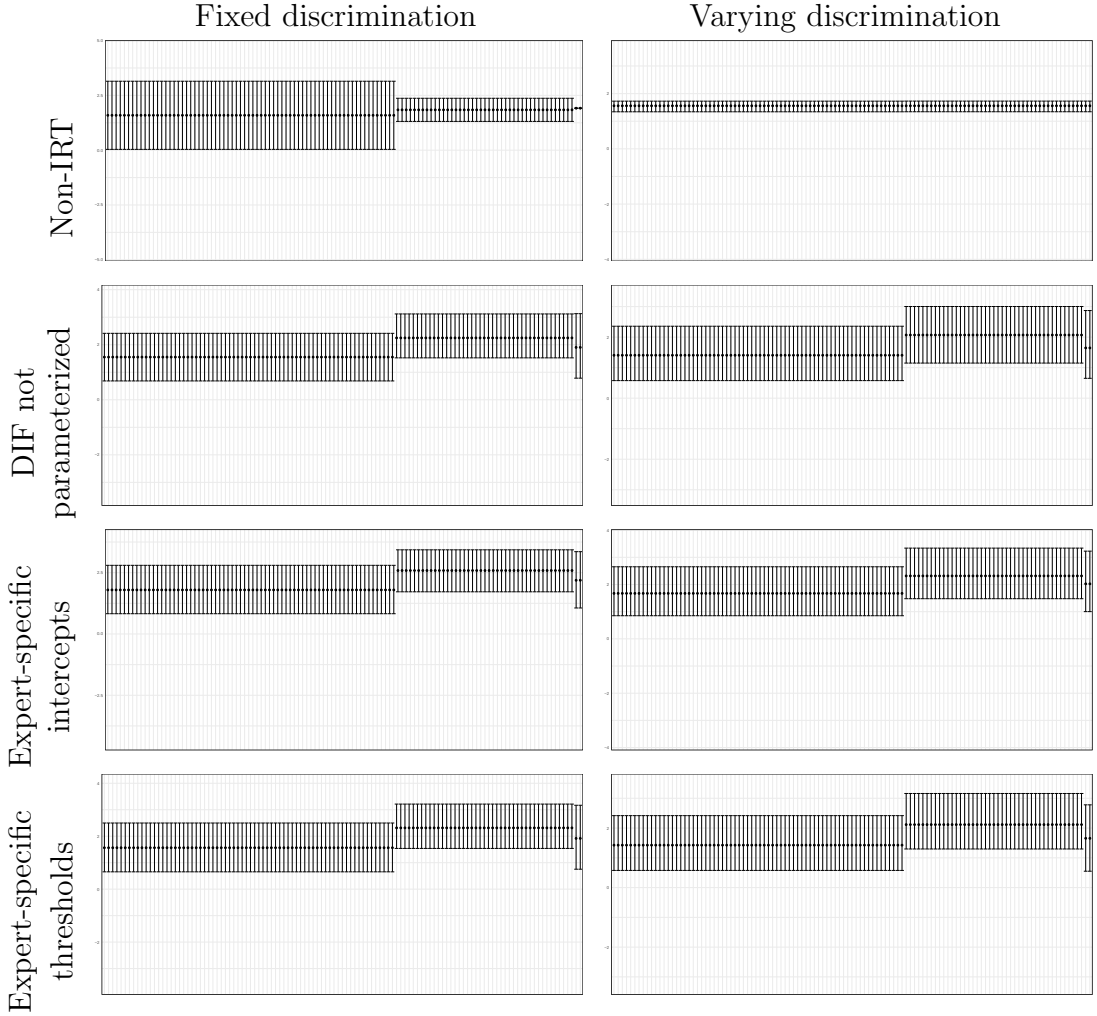
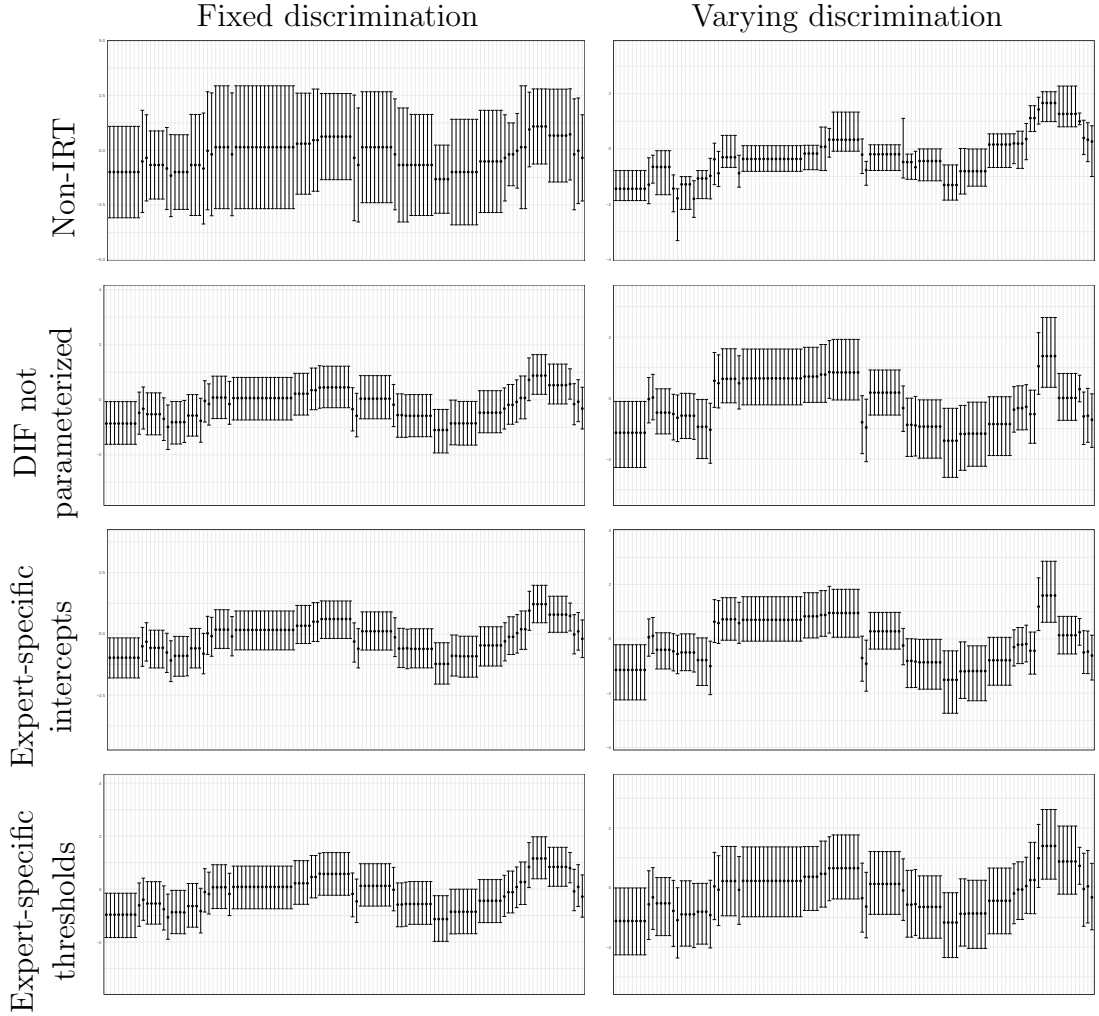


Figure E.3: Different IRT models of freedom from political killings in Turkey



F Simulation algorithm

1. Estimate true value ξ for country-year ct by taking the mean of expert codings for each country-year, then normalizing across country-years.

2. Simulate reliability and agreement values

   - Simulate reliability β for expert r
     - No variation $\beta_r = \beta = 1$
     - Medium variation: $\beta_r \sim \mathcal{N}(1, 0.5)$
     - High variation: $\beta_r \sim \mathcal{N}(1, 1)$
   - Simulate expert agreement parameters
     - Perfect agreement
       - $\tau_{r;1,2,3,4} = \gamma_{1,2,3,4} = (-0.88, -0.31, 0.14, 0.83)$
       - $\kappa_r = \kappa = 0$
     - Simulate intercept parameter κ for expert r
       (a) Simulate κ for main country-coded c_r
       - Medium variation: $\kappa^{c_r} \sim \mathcal{N}(0, 0.5)$
       - High variation: $\kappa^{c_r} \sim \mathcal{N}(0, 1)$

       (b) Simulate κ for expert r
       - Medium variation: $\kappa_r \sim \mathcal{N}(\kappa^{c_r}, 0.5)$
       - High variation: $\kappa_r \sim \mathcal{N}(\kappa^{c_r}, 1)$

       (c) Create expert thresholds with formula $\tau_{r,k} = \gamma_k + \kappa_r$
     - Simulate threshold parameters τ for expert r and threshold k, $\kappa = 0$
       (a) Simulate τ for main country-coded c_r
       - Medium variation: $\tau_k^{c_r} \sim \mathcal{N}(\gamma_k, 0.25)$
       - High variation: $\tau_k^{c_r} \sim \mathcal{N}(\gamma_k, 1)$

       (b) Order $\tau_k^{c_r}$

       (c) Simulate τ for expert r
       - Medium variation: $\tau_{r,k} \sim \mathcal{N}(\tau_k^{c_r}, 0.25)$
       - High variation: $\tau_{r,k} \sim \mathcal{N}(\tau_k^{c_r}, 1)$

       (d) Order $\tau_{r,k}$
     - Simulate truncated threshold parameters τ for expert r and threshold k, $\kappa = 0$
       (a) Assign main country-coded c_r indicator $\zeta^{c_r} \sim Bernoulli(0.5)$ for positive or negative truncation

       (b) Simulate τ for main country-coded c_r
       - Medium variation: $\tau_k^{c_r} \sim \mathcal{N}(\gamma_k, 0.25)$
         - If $\zeta^{c_r} = 1$, $\min(\tau_{r,k}) = \gamma_k$
         - If $\zeta^{c_r} = 0$, $\max(\tau_{r,k}) = \gamma_k$
       - High variation: $\tau_k^{c_r} \sim \mathcal{N}(\gamma_k, 1)$, truncated as with medium variation

(c) Order $\tau_k^{c_r}$

(d) Assign expert r indicator $\zeta_r \sim Bernoulli(0.5)$ for positive or negative truncation

(e) Simulate τ for expert r

* Medium variation: $\tau_{r,k} \sim \mathcal{N}(\tau_k^{c_r}, 0.25)$
  - If $\zeta_r = 1$, $\min(\tau_k) = \tau_k^{c_r}$
  - If $\zeta_r = 0$, $\max(\tau_k) = \tau_k^{c_r}$
* High variation: $\tau_{r,k} \sim \mathcal{N}(\tau_k^{c_r}, 1)$, truncated as with medium variation

(f) Order $\tau_{r,k}$

3. Create perceived latent values λ for expert r and country year ct with equation $\lambda_{rct} = \beta_r \xi_{ct}$

4. Observed score $y_{rct} \sim Categorical(p_{krct})$, where $p_{krct} = \phi(\tau_{r,k} - \lambda_{rct}) - \phi(\tau_{r,k-1} - \lambda_{rct})$ and ϕ is the CDF of a normal distribution

   - Simulate observed scores for all permutations of β (no variation, medium variation, and high variation) and τ (perfect agreement, medium and high intercept variation, medium and high threshold variance, and medium and high truncated threshold variance).
   - Total number of permutations of simulated data: $3 \times (1 + 2 + 2 + 2) = 21$

5. Repeat thrice to create three unique data sets with 21 combinations

G Additional model fit figures

G.1 Additional MSE figures

G.1.1 Models with saturated data, all possible bridging

Figure G.1: MSE estimates across simulations with no DIF, using simulated data with all possible bridging.

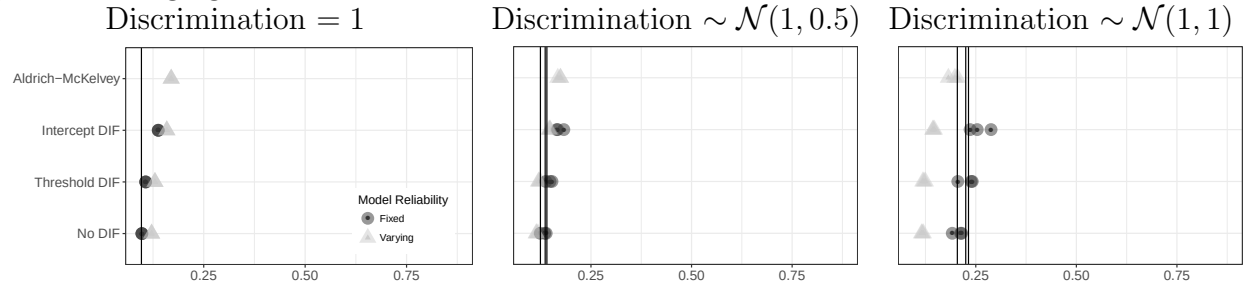
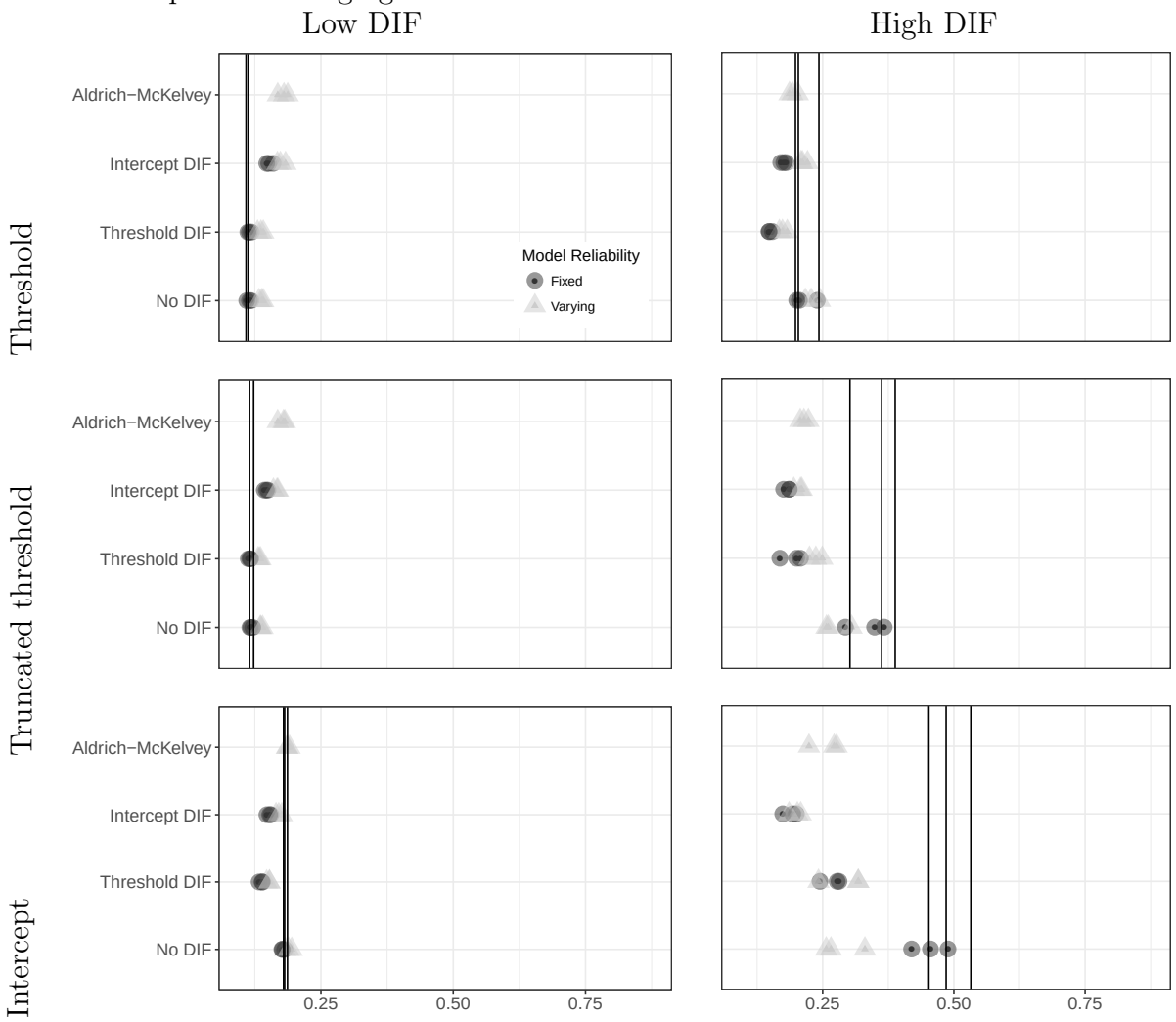
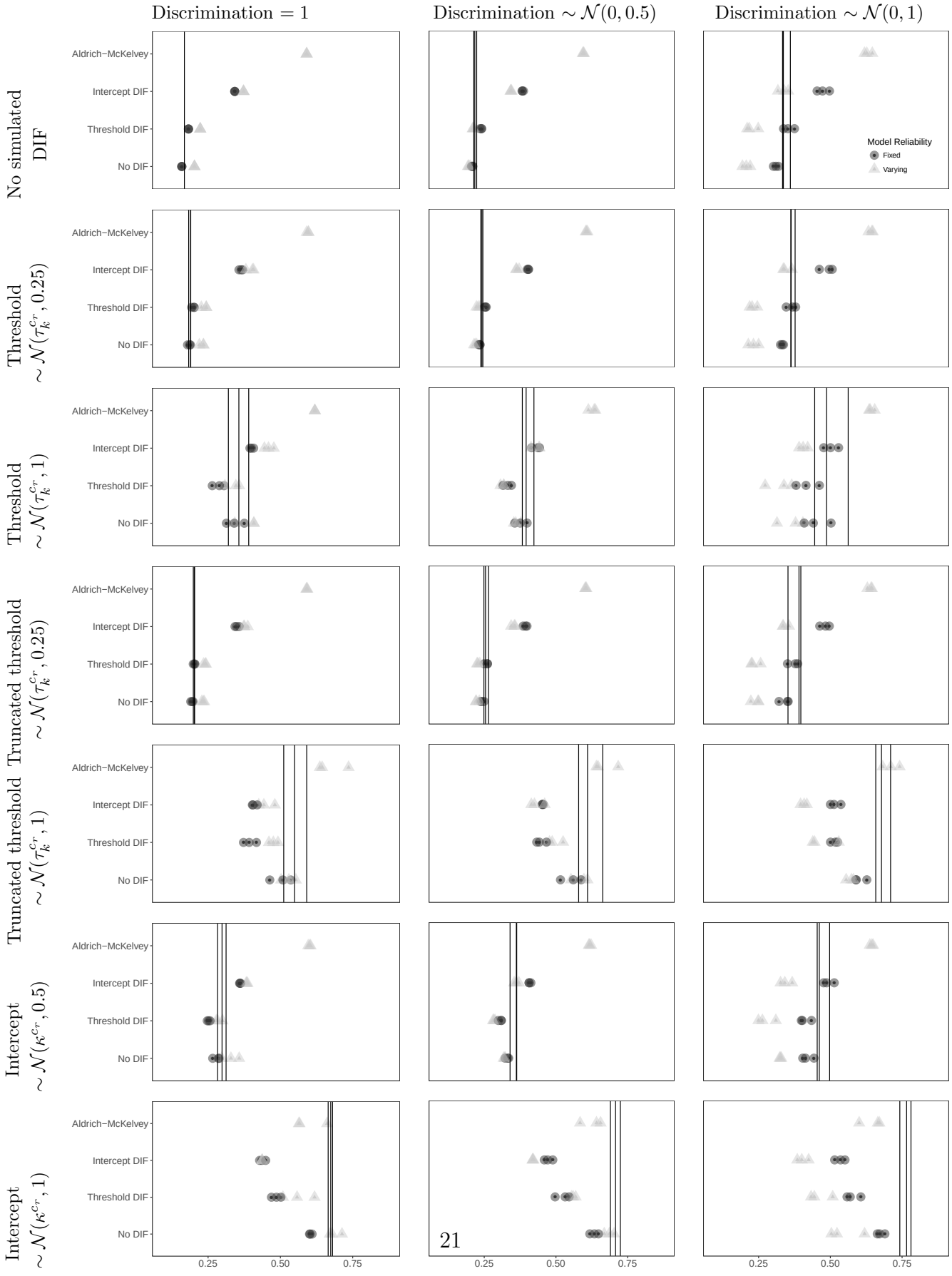


Figure G.2: MSE estimates across simulations with fixed discrimination, using saturated data with all possible bridging.



G.1.2 Models with saturated data, no bridging

Figure G.3: Saturated data with no bridging



G.2 Pearson correlation estimates across simulations

Figure G.4: Data with V–Dem structure

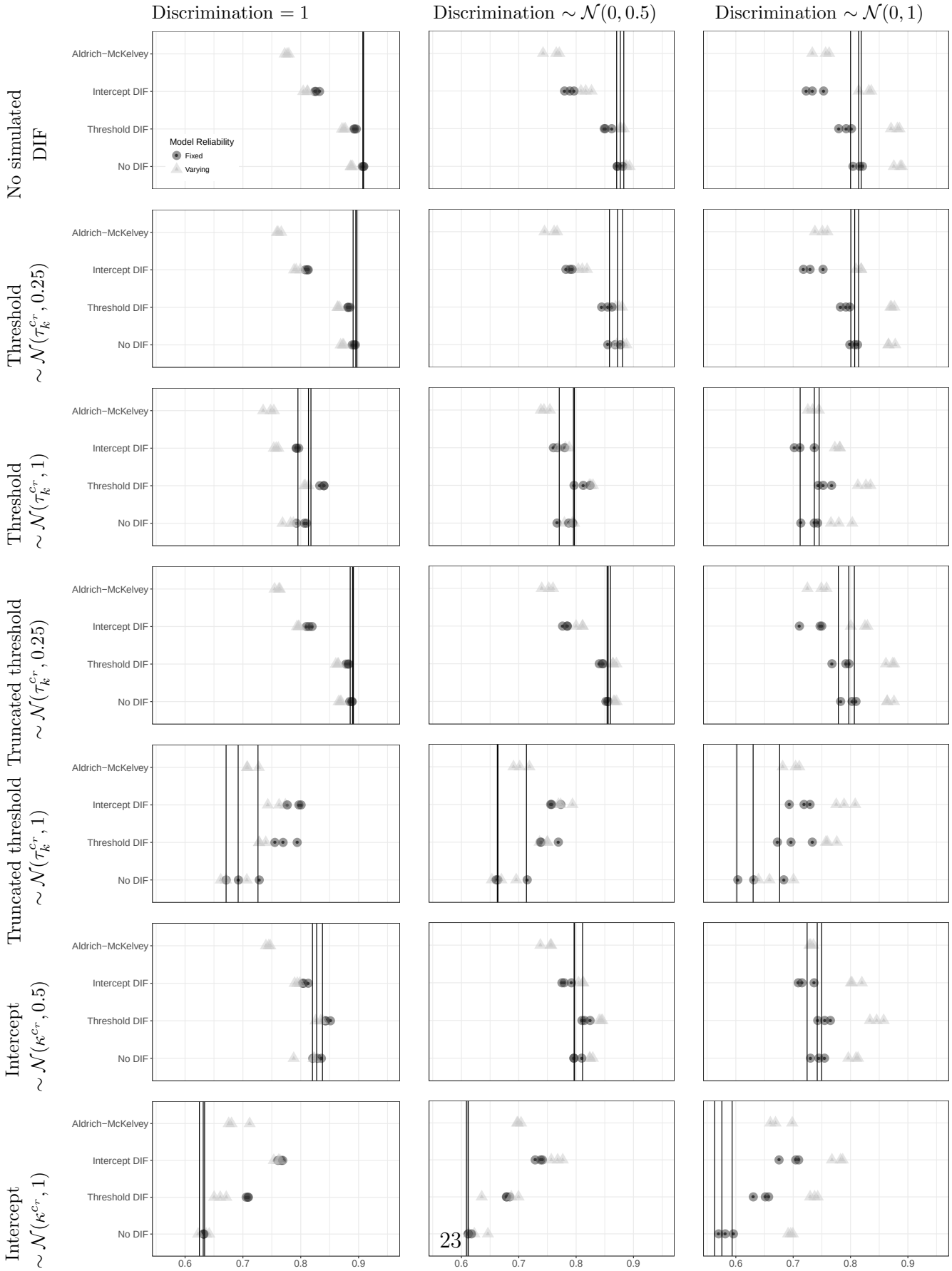


Figure G.5: Saturated data with no bridging

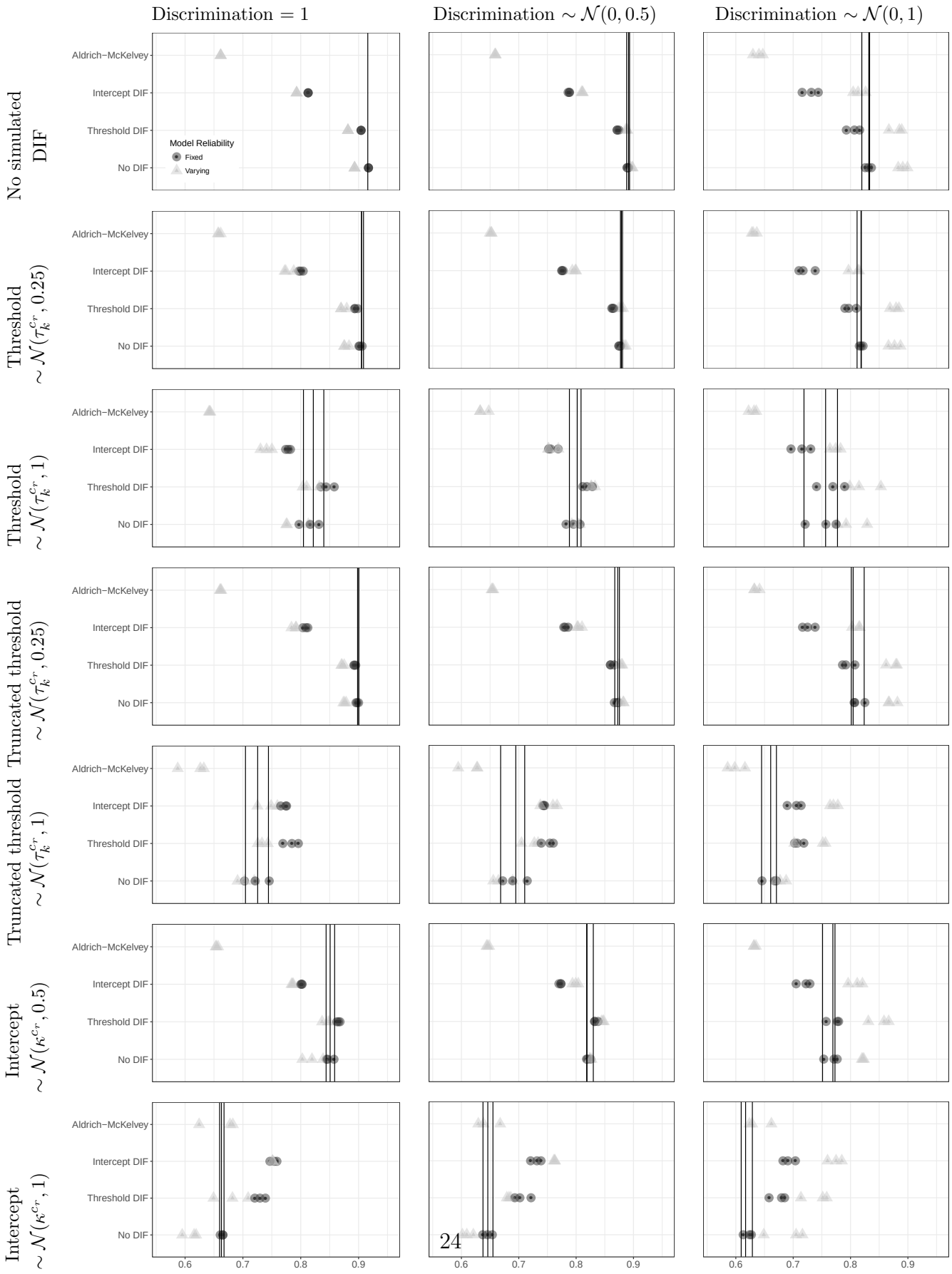
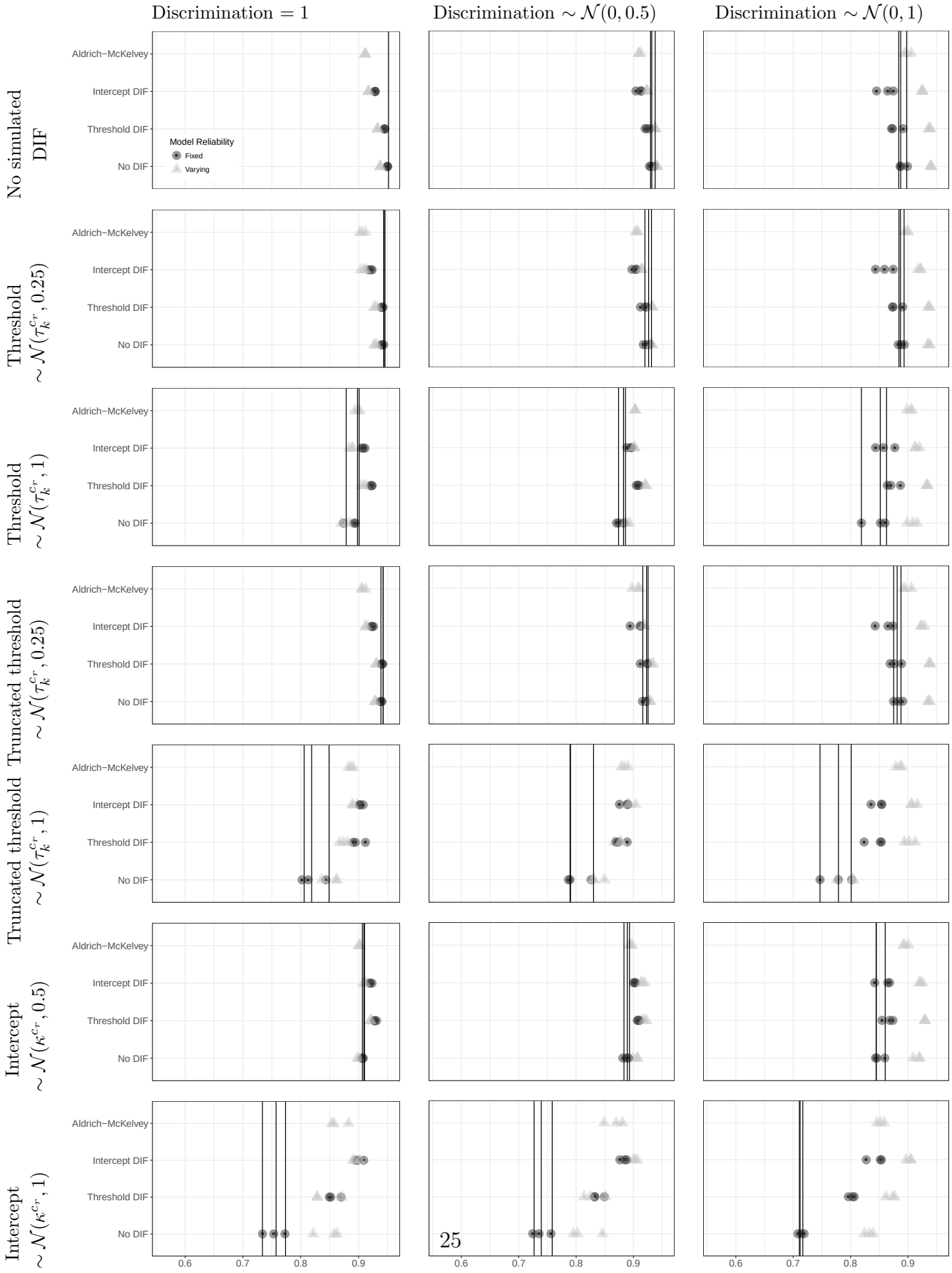


Figure G.6: Saturated data with all possible bridging



G.3 Kendall correlation estimates across simulations

Figure G.7: Data with V–Dem structure

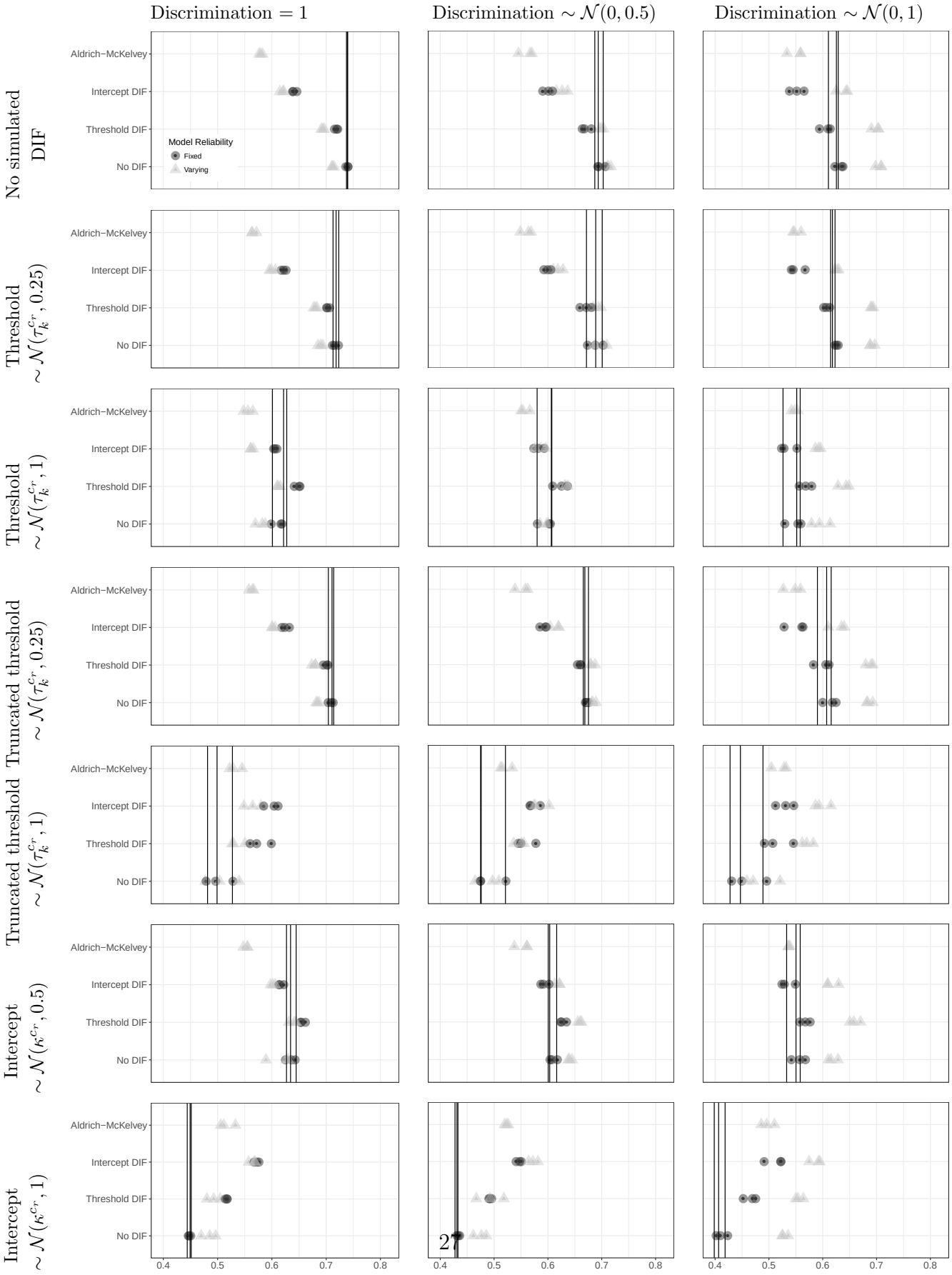


Figure G.8: Saturated data with no bridging

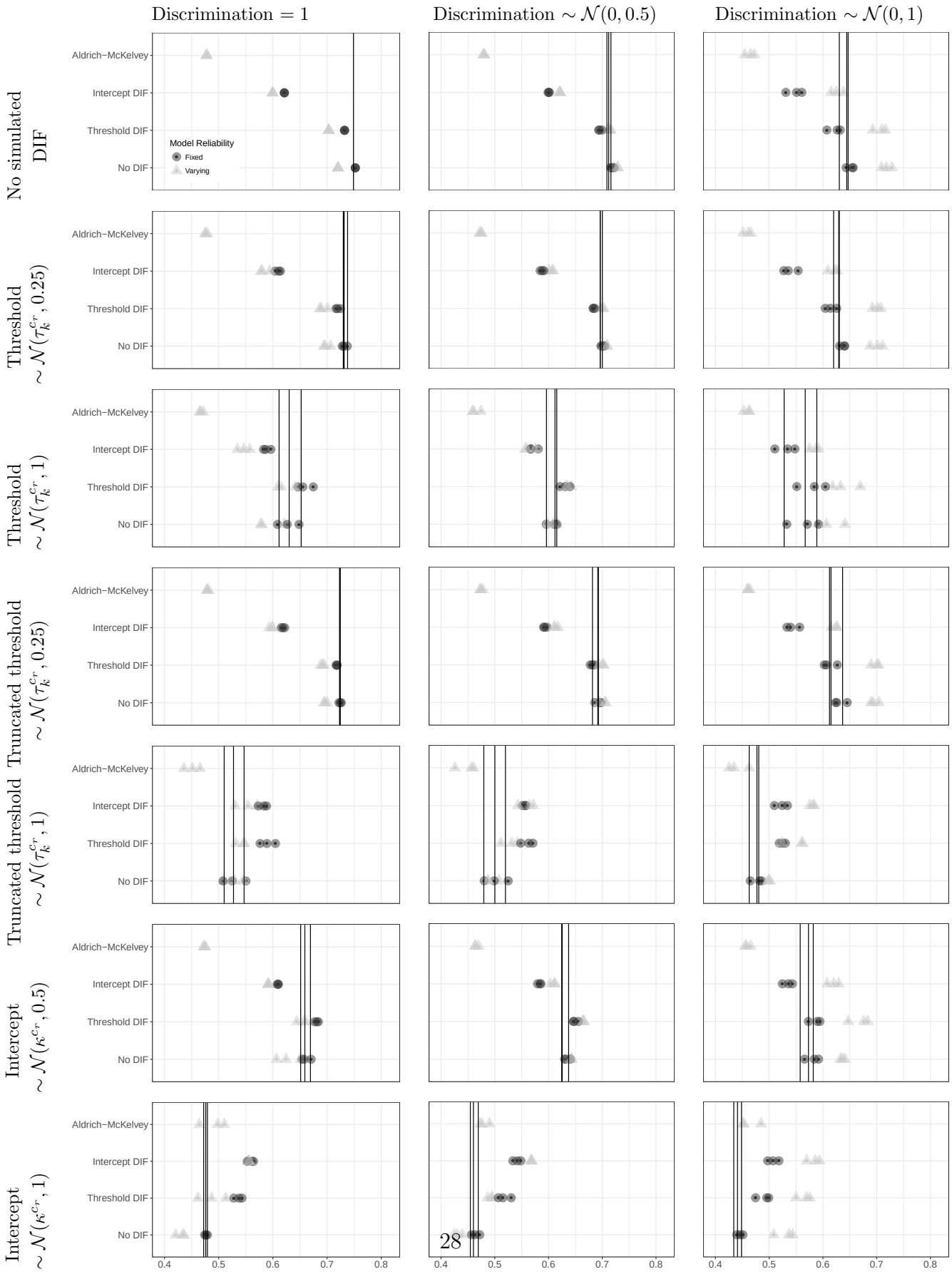
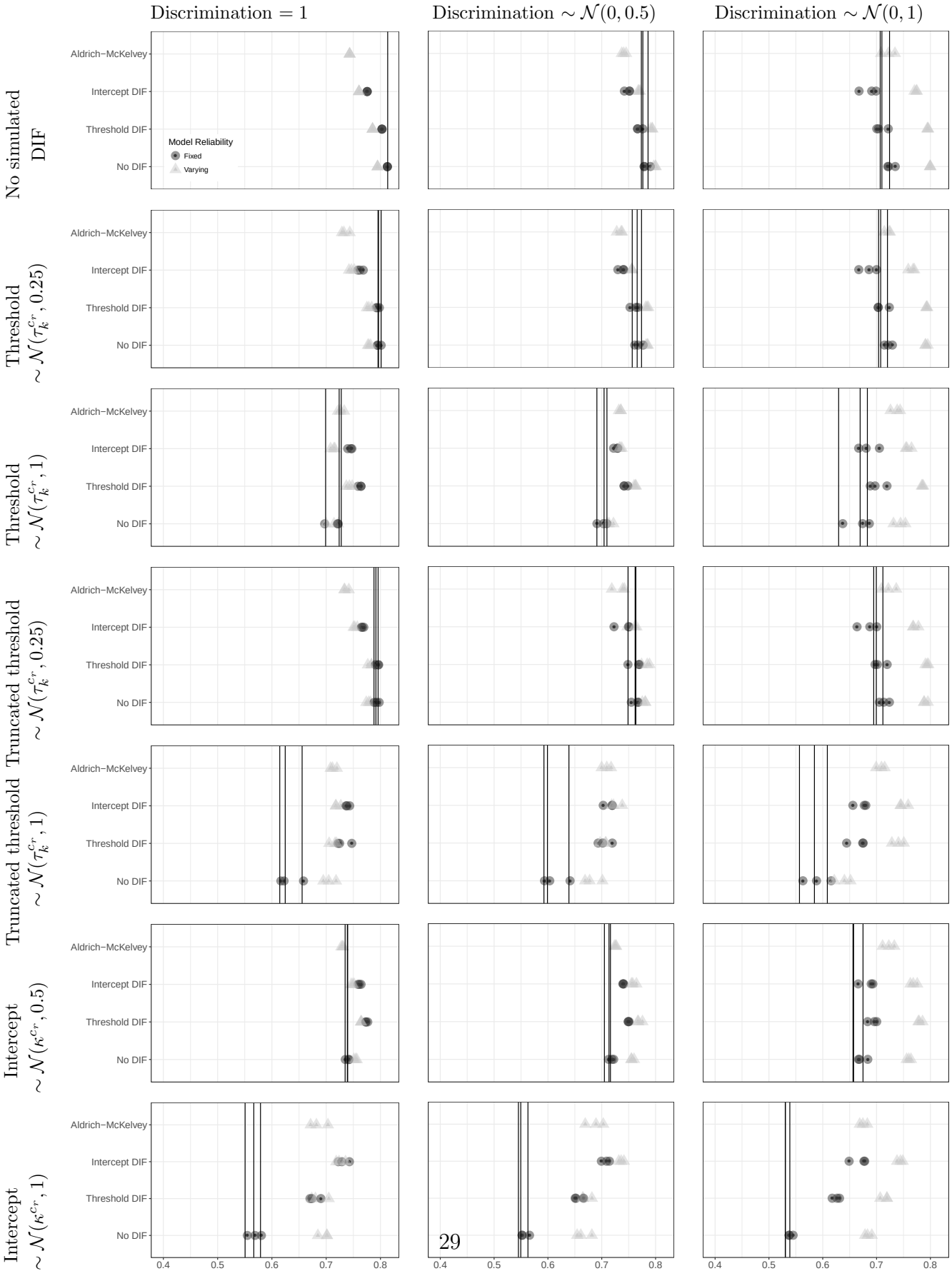


Figure G.9: Saturated data with all possible bridging



G.4 HPD estimates across simulations

Figure G.10: Data with V–Dem structure

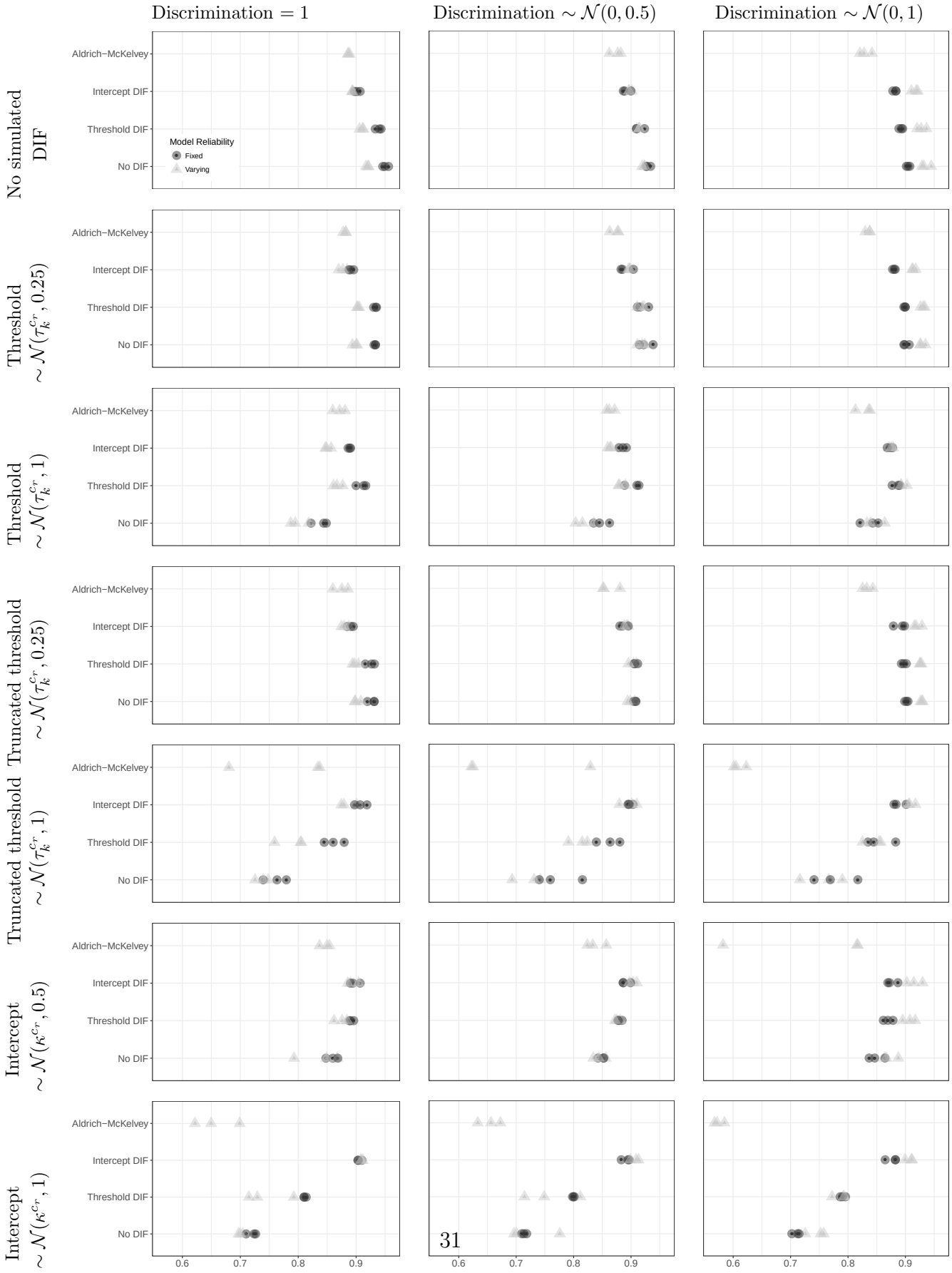


Figure G.11: Saturated data with no bridging

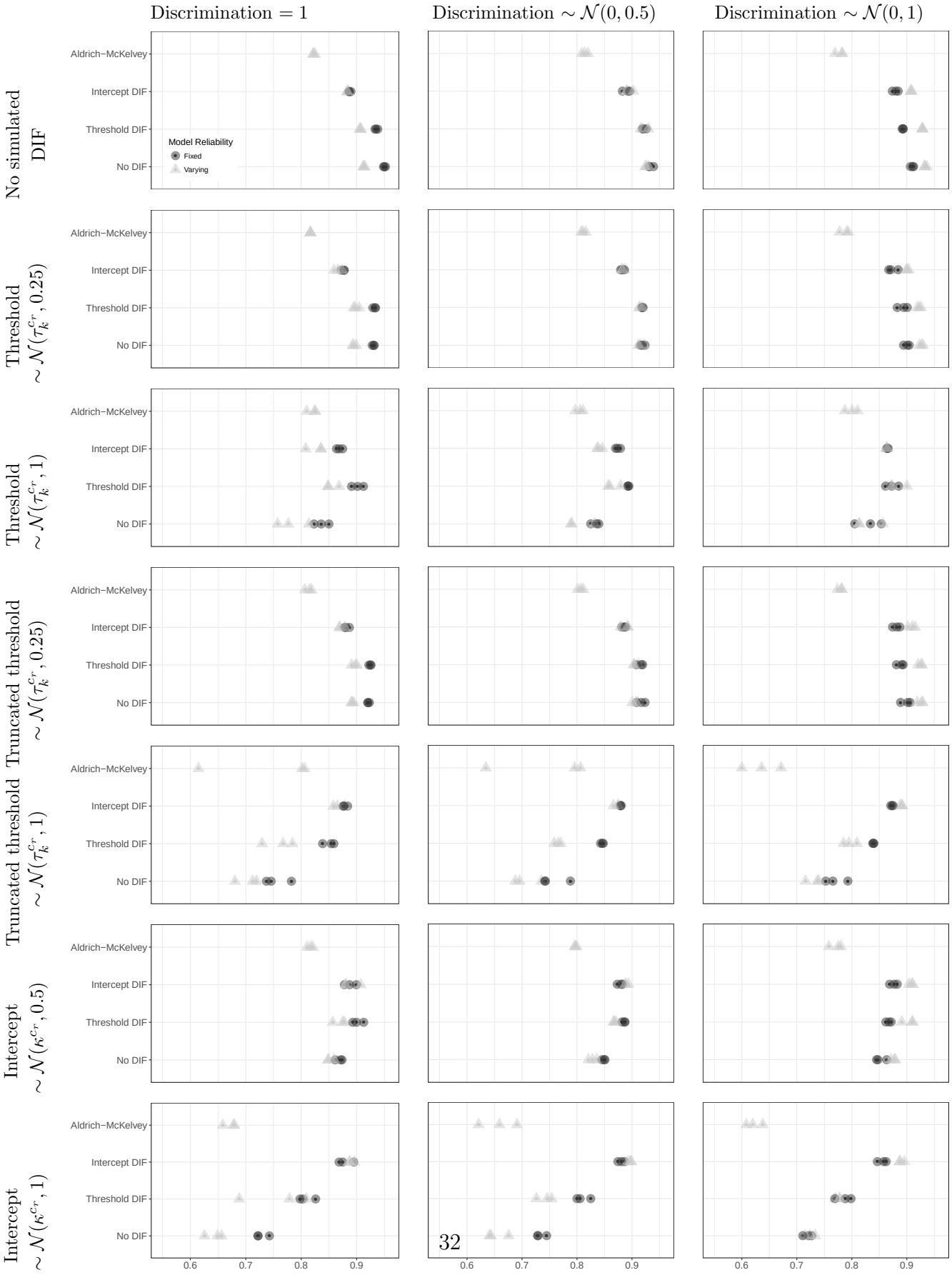
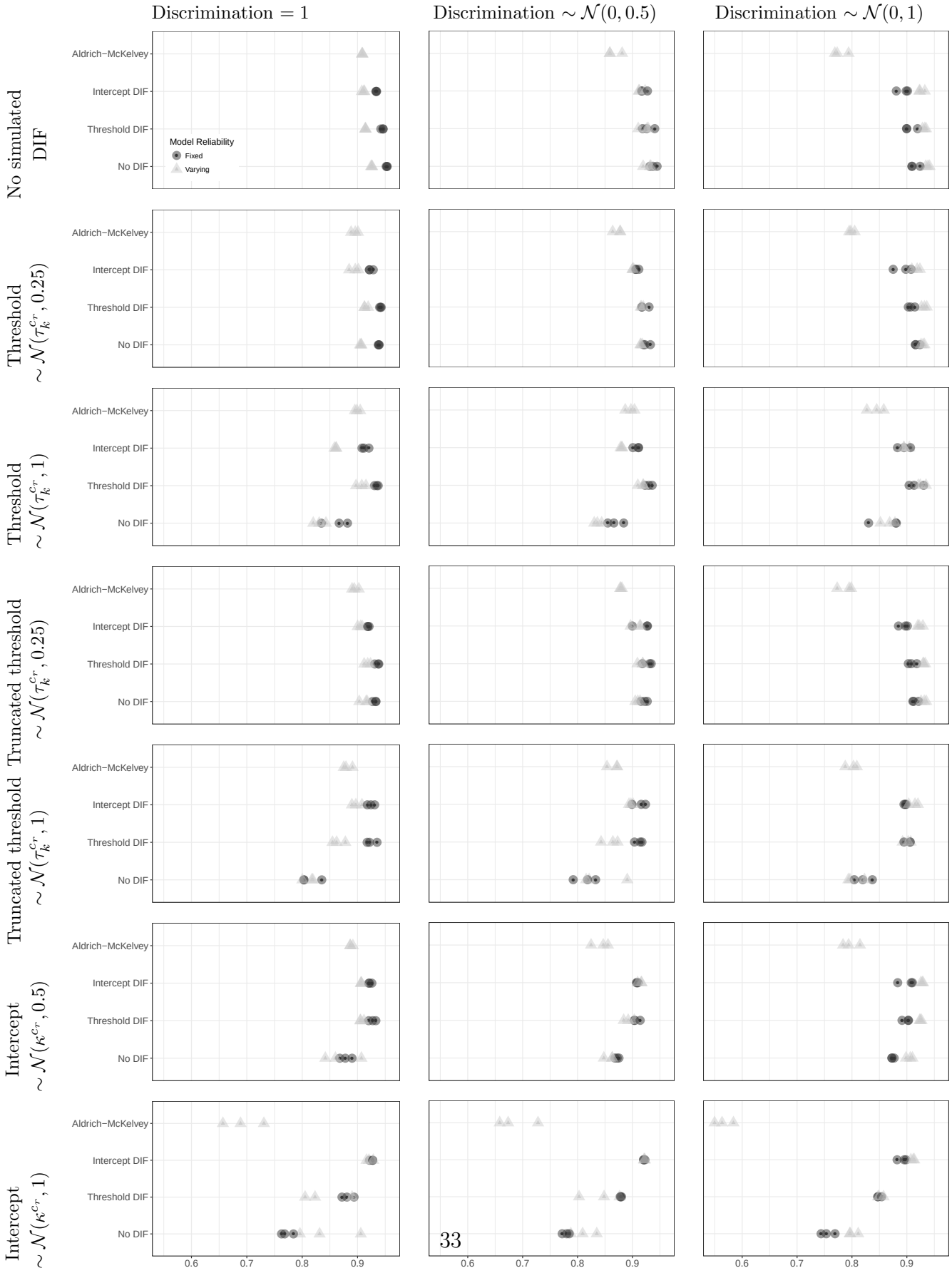


Figure G.12: Saturated data with all possible bridging



H Comparison of Cauchy and uniform prior distributions, V–Dem bridging structure

Table H.1: Results from analyses of simulated data without DIF and different levels of simulated reliability

	Simulated reliability= 1				Simulated reliability$\sim N(1,1)$			
	MSE	Pearson	Kendall	HPD	MSE	Pearson	Kendall	HPD
No DIF in model								
Model reliability= 1								
Uniform	.17 (.17, .18)	.91 (.91, .91)	.74 (.74, .74)	.95 (.95, .96)	.34 (.33, .36)	.82 (.80, .82)	.64 (.62, .64)	.90 (.90, .91)
Cauchy	.17	.91	.74	.95	.33	.82	.64	.91
Varying reliability in model								
Uniform	.21 (.21, .22)	.89 (.89, .89)	.71 (.71, .71)	.92 (.92, .92)	.21 (.22, .23)	.89 (.88, .89)	.71 (.70, .71)	.93 (.93, .94)
Cauchy	.22	.89	.71	.92	.21	.89	.71	.93
Hierarchical intercept DIF in model								
Model reliability= 1								
Uniform	.32 (.31, .32)	.83, (.82, .83)	.64 (.64, .65)	.90 (.90, .91)	.47 (.44, .49)	.73 (.72, .75)	.55 (.54, .57)	.88 (.88, .88)
Cauchy	.32	.82	.64	.90	.47	.73	.55	.88
Varying reliability in model								
Uniform	.34 (.34, .35)	.81 (.80, .81)	.62 (.61, .62)	.89 (.89, .89)	.31 (.30, .34)	.83 (.81, .84)	.64 (.62, .65)	.92 (.91, .92)
Cauchy	.35	.80	.61	.90	.31	.83	.64	.92

Table H.2: Results from analyses of simulated data with threshold variance $\sim N(\tau_k^{c_r}, 1)$ and different levels of simulated reliability

	Simulated reliability= 1				Simulated reliability$\sim N(1,1)$			
	MSE	Pearson	Kendall	HPD	MSE	Pearson	Kendall	HPD
No DIF in model								
Model reliability= 1								
Uniform	.36 (.35, .38)	.81 (.79, .81)	.62 (.60, .62)	.84 (.82, .85)	.47 (.46, .51)	.74 (.71, .74)	.55 (.53, .56)	.84 (.82, .85)
Cauchy	.38	.79	.60	.82	.51	.71	.53	.82
Varying reliability in model								
Uniform	.39 (.38, .42)	.78 (.77, .79)	.58 (.57, .59)	.79 (.79, .82)	.40 (.36, .42)	.78 (.77, .80)	.59 (.58, .61)	.84 (.83, .86)
Cauchy	.42	.77	.57	.79	.43	.76	.58	.84
Hierarchical intercept DIF in model								
Model reliability= 1								
Uniform	.37 (.37, .36)	.79 (.79, .80)	.61 (.60, .61)	.89 (.89, .89)	.50 (.47, .52)	.71 (.70, .74)	.53 (.52, .55)	.87 (.87, .88)
Cauchy	.38	.79	.60	.89	.50	.71	.53	.87
Varying reliability in model								
Uniform	.43 (.42, .44)	.76 (.75, .76)	.56 (.56, .57)	.85 (.85, .86)	.39 (.39, .41)	.78 (.77, .78)	.59 (.59, .60)	.88 (.87, .88)
Cauchy	.44	.75	.56	.85	.41	.77	.59	.87

Table H.3: Results from analyses of simulated data with truncated threshold variance $\sim N(\tau_k^{cr}, 1)$ and different levels of simulated reliability

	Simulated reliability= 1				Simulated reliability$\sim N(1,1)$			
	MSE	Pearson	Kendall	HPD	MSE	Pearson	Kendall	HPD
No DIF in model								
Model reliability= 1								
Uniform	.56 (.49, .59)	.69 (.67, .73)	.50 (.48, .53)	.76 (.74, .78)	.65 (.56, .69)	.63 (.60, .68)	.45 (.43, .50)	.77 (.74, .82)
Cauchy	.49	.73	.53	.78	.56	.68	.50	.81
Varying reliability in model								
Uniform	.58 (.52*, .59)	.67 (.66*, .71)	.50 (.48*, .54)	.74* (.73, .75)	.60 (.52*, .62)	.66 (.64*, .70)	.47* (.46, .52)	.77 (.72, .79*)
Cauchy	.52	.71	.54	.75	.52	.70	.52	.77
Hierarchical intercept DIF in model								
Model reliability= 1								
Uniform	.37 (.36, .40)	.80 (.78, .80)	.61 (.58, .60)	.91 (.90, .92)	.49 (.47, .53)	.72 (.69, .73)	.53 (.51, .55)	.88 (.88, .90)
Cauchy	.36	.80	.61	.92	.47	.73	.54	.90
Varying reliability in model								
Uniform	.42 (.40, .45)	.76 (.74, .77)	.56 (.55, .58)	.88 (.87, .90)	.38 (.35, .40)	.79 (.78, .81)	.59 (.59, .62)	.91 (.91, .92)
Cauchy	.40	.77	.58	.90	.35	.81	.62	.92

*: Represents models that did not converge after 10,000 iterations with eight chains. Bold text represents model with highest correlation.

Table H.4: Results from analyses of simulated data with intercept variance $\sim N(\kappa^{c_r}, 1)$ and different levels of simulated reliability

	Simulated reliability= 1				Simulated reliability$\sim N(1,1)$			
	MSE	Pearson	Kendall	HPD	MSE	Pearson	Kendall	HPD
No DIF in model								
Model reliability= 1								
Uniform	.66 (.65, .66)	.63 (.63, .63)	.45 (.45, .45)	.72 (.71, .73)	.74 (.71, .76)	.58 (.57, .60)	.41 (.40, .42)	.71 (.70, .71)
Cauchy	.66	.63	.45	.72	.71	.60	.42	.71
Varying reliability in model								
Uniform	.63 (.62, .66)	.64 (.62, .64)	.49 (.47, .50)	.70 (.70, .70)	.54 (.53, .55)	.70 (.69, .70)	.53 (.52, .54)	.75 (.73, .76)
Cauchy	.63	.64	.50	.71	.53	.70	.54	.76
Hierarchical intercept DIF in model								
Model reliability= 1								
Uniform	.41 (.41, .42)	.77 (.76, .77)	.57 (.57, .58)	.90 (.90, .91)	.51 (.50, .56)	.70 (.68, .71)	.52 (.49, .52)	.88 (.86, .88)
Cauchy	.41	.77	.57	.91	.51	.70	.52	.89
Varying reliability in model								
Uniform	.42 (.42, .43)	.76 (.75, .76)	.57 (.56, .57)	.91 (.91, .91)	.39 (.38, .41)	.78 (.77, .79)	.59 (.57, .59)	.91 (.90, .91)
Cauchy	.43	.75	.56	.91	.38	.79	.59	.91