Building Bayesian Credible Intervals for Classification Accuracy Analysis Indices

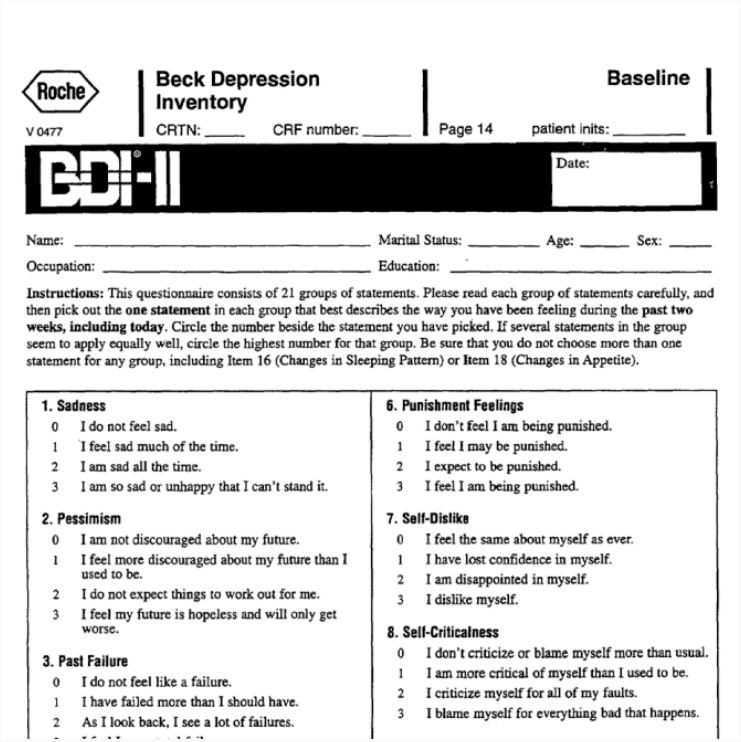
Meltem Ozcan

April 26, 2022

PSYC 573 Bayesian Data Analysis

Introduction: Psychometric tests

- Unobservable (latent) vs. observable (manifest) variables
  - e.g. true level of depression vs. scores on a test of depression
- Test scores as an approximation to the 'true' score
- Can be used to classify/select individuals into categories:
  - Total scores on the BDI-II -> minimal, mild, moderate, severe depression
  - Aptitude test for potential employer -> accepted, rejected
- High stakes decision-making



Roche | Beck Depression Inventory | Baseline

V 0477 CRTN: ____ CRF number: ____ Page 14 patient inits: ____

BDI-II Date:

Name: ____ Marital Status: ____ Age: ____ Sex: ____

Occupation: ____ Education: ____

Instructions: This questionnaire consists of 21 groups of statements. Please read each group of statements carefully, and then pick out the **one statement** in each group that best describes the way you have been feeling during the **past two weeks, including today**. Circle the number beside the statement you have picked. If several statements in the group seem to apply equally well, circle the highest number for that group. Be sure that you do not choose more than one statement for any group, including Item 16 (Changes in Sleeping Pattern) or Item 18 (Changes in Appetite).

1. Sadness
0 I do not feel sad.
1 I feel sad much of the time.
2 I am sad all the time.
3 I am so sad or unhappy that I can't stand it.

2. Pessimism
0 I am not discouraged about my future.
1 I feel more discouraged about my future than I used to be.
2 I do not expect things to work out for me.
3 I feel my future is hopeless and will only get worse.

3. Past Failure
0 I do not feel like a failure.
1 I have failed more than I should have.
2 As I look back, I see a lot of failures.

6. Punishment Feelings
0 I don't feel I am being punished.
1 I feel I may be punished.
2 I expect to be punished.
3 I feel I am being punished.

7. Self-Dislike
0 I feel the same about myself as ever.
1 I have lost confidence in myself.
2 I am disappointed in myself.
3 I dislike myself.

8. Self-Criticalness
0 I don't criticize or blame myself more than usual.
1 I am more critical of myself than I used to be.
2 I criticize myself for all of my faults.
3 I blame myself for everything bad that happens.

Figure 1: Excerpt of the Beck Depression Inventory (BDI-II)

The Common Factor Model (Thurstone, 1947)

- True standing on unobserved construct governs P(observed responses) through a system of linear equations
- For a one-factor model,

$$\boldsymbol{x}_i = \boldsymbol{\nu} + \boldsymbol{\lambda}\eta_i + \boldsymbol{\epsilon}_i$$

$\boldsymbol{x}_i$: observed item scores ($J \times 1$)
η_i: latent factor score
J: number of observed variables
i: individual ($i = 1, \ldots, N$)

α: latent factor mean
ψ: latent factor variance
$\boldsymbol{\lambda}$: factor loadings (J x 1)
$\boldsymbol{\nu}$: measurement intercepts (J x 1)
ϵ_i: unique factor variables with E(ϵ)=0, Var(ϵ)=θ

- Assuming $\text{Cov}(\boldsymbol{\epsilon}, \eta) = 0$, observed variables are distributed with $\text{E}(\boldsymbol{x}) = \boldsymbol{\nu} + \boldsymbol{\lambda\alpha}$ and $\boldsymbol{\Sigma} = \boldsymbol{\lambda\theta\lambda}' + \boldsymbol{\Psi}$

(Levy & Mislevy, 2016; Meredith & Teresi, 2006)

Measurement Invariance (MI)

- Equivalence of measurement operations across groups and conditions
  - ◇ Mathematically, $P(\boldsymbol{X}|\boldsymbol{\eta}, G = g) = P(\boldsymbol{X}|\boldsymbol{\eta}), \forall g$ (Mellenbergh, 1989)

    η: latent factor scores
    X: observed scores
    g: group membership
- Factorial invariance $\equiv$ measurement invariance (Horn & McArdle, 1992)
  - ◇ equal intercepts
  - ◇ equal loadings
  - ◇ equal unique factor variances
- Partial invariance (Byrne et al., 1989)

Selection Accuracy Analysis Framework (Millsap & Kwok, 2004)

Evaluate the practical impact of partial invariance on selection outcomes

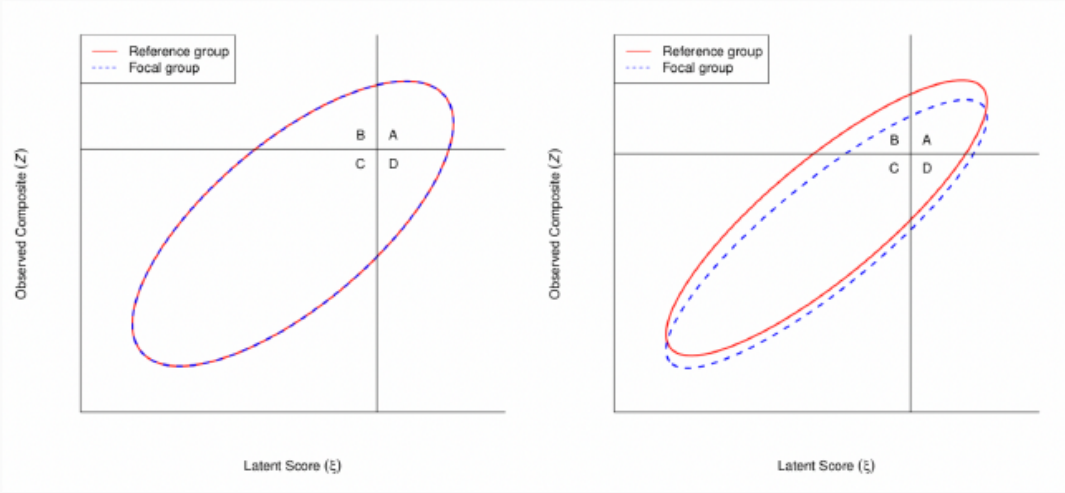


Figure 2: Left: strict invariance. Right: partial invariance

True value \ Classification outcome	+	-
+	True Positive (A)	False Negative (D)
-	False Positive (B)	True Negative (C)

$$\text{Proportion Selected (PS)} = P(TP) + P(FP)$$

$$\text{Success Ratio (SR)} = \frac{P(TP)}{P(TP) + P(FP)}$$

$$\text{Sensitivity (SE)} = \frac{P(TP)}{P(TP) + P(FN)}$$

$$\text{Specificity (SP)} = \frac{P(TN)}{P(TN) + P(FP)}$$

Method

- Step 1: Fit a Bayesian Confirmatory Factor Analysis (CFA) model using R package *blavaan* (Merkle et al., 2021)
- Step 2: Extract the posterior distribution of the latent variable (η) conditioned on the observed variables
- Step 3: Compute the posterior expected distribution of the observed variables conditioned on the sampled latent variables
- Step 4: Compute the scale sums (Z) for each iteration
- Step 5: Compute cut-off scores using the full sample
- Step 6: Compute the posterior distribution of selection accuracy indices for the reference and focal groups
- Step 7: Compute the posterior distribution of the effect size of the difference between selection accuracy indices for reference vs. focal groups
- Step 8: Compute CI based on the posterior distribution

Data and Participants

- Data collected in 1994-1996 for the development of the International Personality Item Pool (Donnellan et al., 2006)
- Measurement invariance of Mini-IPIP across gender (Ock et al., 2020)
- 20 items measured on a 1-5 Likert scale, 5 facets:
  - Agreeableness
  - Conscientiousness
  - Extraversion
  - Neuroticism
  - Openness to Experience
- 564 participants
  - $n_{male} = 239$, $n_{female} = 325$ - F coded as "1", M coded as "2"
  - Age: range of 20-85, $M = 51.7$, $SD = 12.5$
  - 97.7% Caucasian

Mini-IPIP factors and corresponding CFA models

a2 "Sympathize with others' feelings."
a5 "Feel others' emotions."
a7 "Am not really interested in others."
a9 "Am not interested in other people's problems."

c3 "Get chores done right away."
c4 "Like order."
c8 "Make a mess of things."
c9 "Often forget to put things back in their proper place."

e1 "Am the life of the party."
e4 "Talk to a lot of different people at parties."
e6 "Don't talk a lot."
e7 "Keep in the background."

n1 "Am nervous most of the time."
n2 "Seldom feel blue."
n6 "Get upset easily."
n8 "Have frequent mood swings."

o2 "Have a vivid imagination."
o8 "Have difficulty understanding abstract ideas."
o9 "Am not interested in abstract ideas."
o10 "Do not have a good imagination."

Agreeableness

$$'A =\sim a2 + a5 + a7 + a9$$

$$a2 \sim\sim a5'$$

Conscientiousness

$$'C =\sim c3 + c4 + c8 + c9'$$

Extraversion

$$'E =\sim e1 + e4 + e6 + e7$$

$$e4 \sim\sim e7'$$

Neuroticism

$$'N =\sim n1 + n2 + n6 + n8'$$

Openness to Experience

$$'O =\sim i2 + i8 + i9 + i10$$

$$i2 \sim\sim i10$$

$$i8 \sim\sim i9'$$

Model

$$\underbrace{p(\boldsymbol{\eta}, \alpha, \psi, \boldsymbol{\nu}, \boldsymbol{\lambda}, \boldsymbol{\theta}|\boldsymbol{X})}_{\text{posterior distribution}} \propto \underbrace{p(\boldsymbol{X}|\boldsymbol{\eta}, \alpha, \psi, \boldsymbol{\nu}\boldsymbol{\lambda}, \boldsymbol{\theta})}_{\text{conditional dist. of observed var.}} \times \underbrace{p(\boldsymbol{\eta}, \alpha, \psi, \boldsymbol{\lambda}, \boldsymbol{\theta})}_{\text{joint dist. of priors}}$$

$$= p(\boldsymbol{X}|\boldsymbol{\eta}, \boldsymbol{\nu}, \boldsymbol{\lambda}, \boldsymbol{\theta}) p(\boldsymbol{\eta}|\alpha, \psi) p(\alpha) p(\psi) p(\boldsymbol{\nu}) p(\lambda) p(\boldsymbol{\theta})$$

$$= \prod_{i=1}^{N} \prod_{j=1}^{j} p(x_{ij}|\eta_i, \nu_i, \lambda_j, \theta_{jj}) p(\eta_i|\alpha, \psi) p(\alpha) p(\psi) p(\nu_j) p(\lambda_j) p(\theta_{jj})$$

$\boldsymbol{X}$: matrix of observed values (N x J)

$\boldsymbol{x}_i$: vector of observed values $\boldsymbol{x}_i = (x_{i1}, ..., x_{iJ})'$

$\boldsymbol{\eta}$: latent variable values (N x 1)

J: # of observed variables

N: # of individuals

α: latent factor mean

ψ: latent factor variance

λ: factor loadings

ν: measurement intercepts

θ: diagonal matrix of Var(ϵ)

Model priors

$$
\begin{aligned}
x_{ij}|\eta_i, \nu_i, \boldsymbol{\lambda}_j, \theta_{jj} &\sim N(\nu_j + \eta_i \boldsymbol{\lambda}_j', \theta_{jj}) \\
\eta_i|\alpha, \psi &\sim N(\alpha, \psi) \\
\alpha &\sim n(0, 10) \\
1/\psi &\sim \text{Gamma}(0.5, 1) \\
\nu_j &\sim N(0, 32) \\
\lambda_j &\sim N(0, 10) \\
\theta_{jj} &\sim \text{Beta}(1, 1)
\end{aligned}
$$

Convergence: Agreeableness model parameters

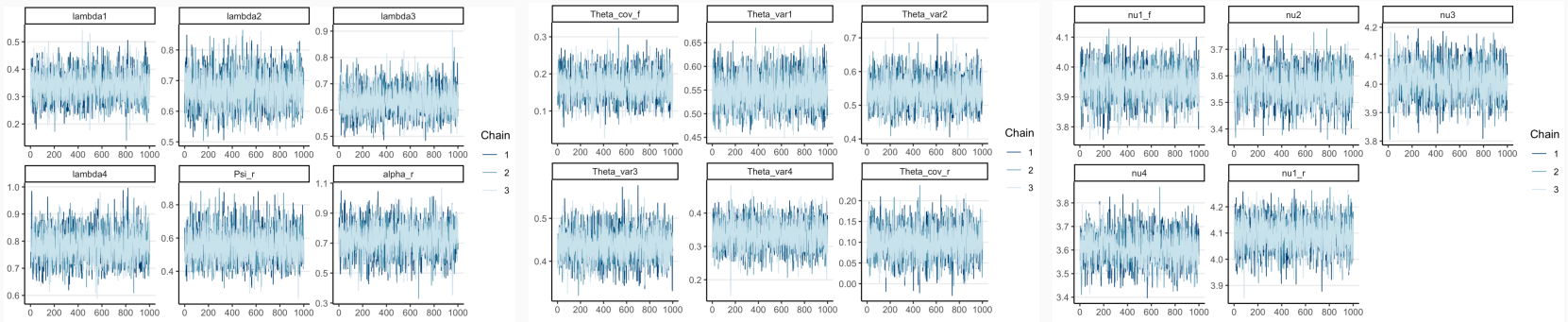


Figure 3: Trace plots for the parameters

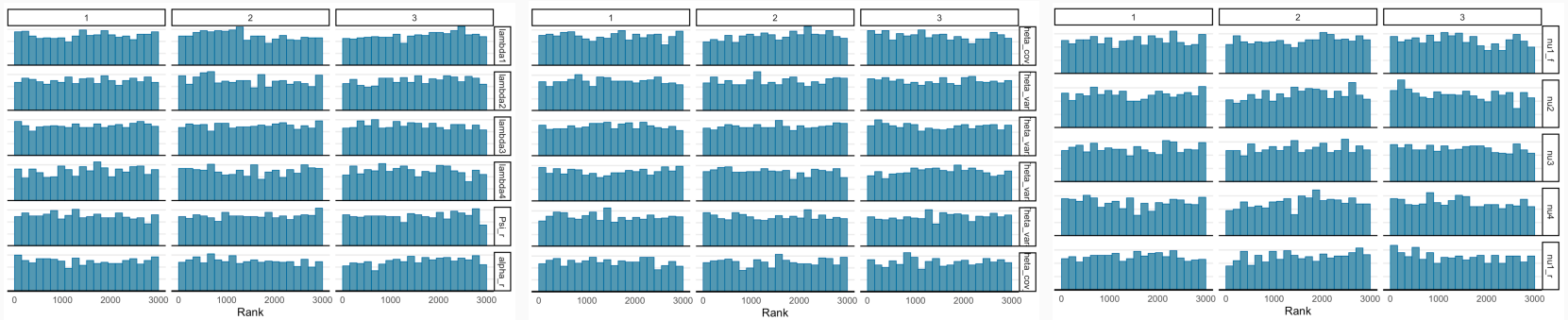


Figure 4: Rank histograms

Posterior summary: Agreeableness

variable	mean	median	sd	mad	q5	q95	rhat	ess_bulk	ess_tail
λ_1	0.34	0.34	0.05	0.05	0.26	0.42	1	2565	2372
λ_2	0.67	0.67	0.06	0.06	0.58	0.76	1	2109	2038
λ_3	0.63	0.63	0.05	0.05	0.55	0.72	1	2598	2252
λ_4	0.78	0.78	0.06	0.06	0.69	0.88	1	2341	2175
θ_1	0.55	0.55	0.03	0.04	0.49	0.61	1	3305	2542
θ_2	0.54	0.54	0.04	0.04	0.47	0.61	1	3192	2395
θ_3	0.44	0.44	0.04	0.04	0.38	0.50	1	3601	2353
θ_4	0.33	0.33	0.04	0.04	0.27	0.40	1	3016	2097
$\theta_{r,c}$	0.17	0.17	0.04	0.04	0.11	0.23	1	3401	1944
$\theta_{f,c}$	0.10	0.10	0.04	0.04	0.04	0.16	1	3366	2211
ν_{1r}	3.95	3.95	0.05	0.05	3.86	4.03	1	2649	2066
ν_{1f}	4.09	4.09	0.06	0.06	4.00	4.19	1	2631	2447
ν_2	3.56	3.56	0.06	0.06	3.46	3.66	1	1979	2235
ν_3	4.01	4.01	0.06	0.06	3.92	4.10	1	1949	1552
ν_4	3.61	3.61	0.06	0.06	3.51	3.71	1	1859	2072
Ψ_f	0.55	0.54	0.09	0.09	0.41	0.72	1	2453	2455
α_f	0.73	0.73	0.10	0.10	0.56	0.89	1	1734	2229

Table 1: Summary of MCMC draws for Agreeableness parameters

Convergence: Conscientiousness

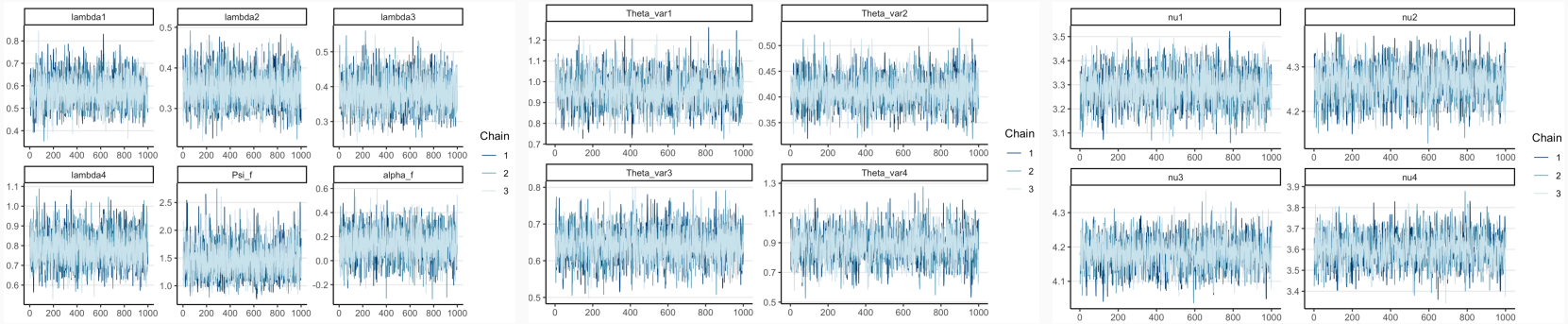


Figure 5: Trace plots

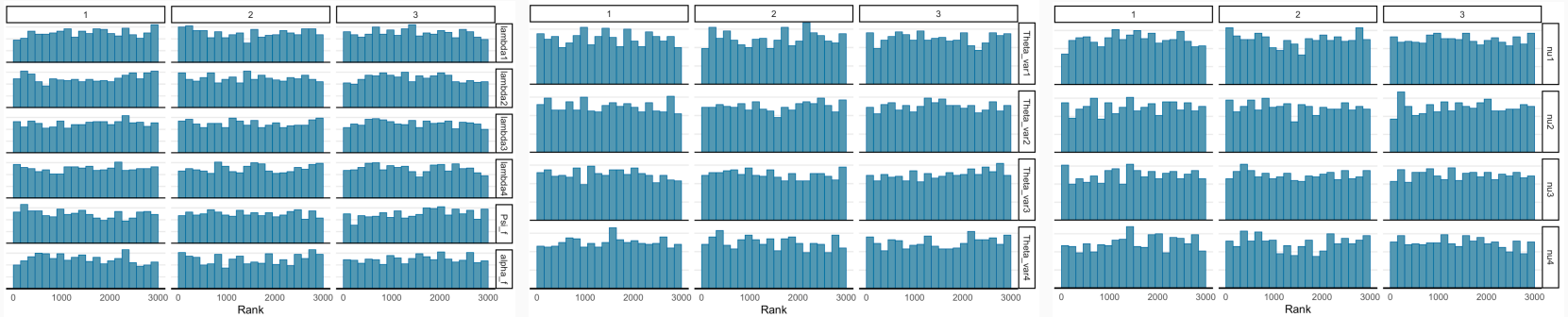


Figure 6: Rank histograms

Convergence: Extraversion

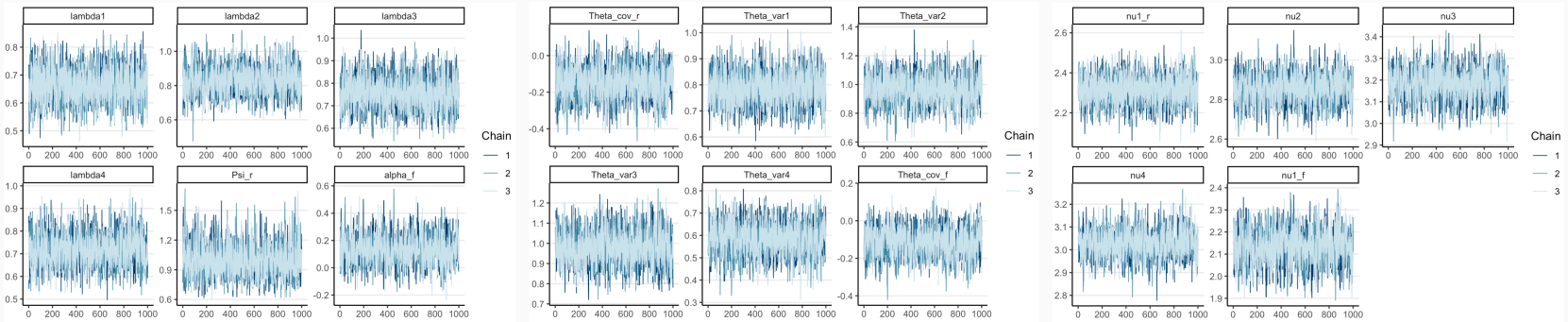


Figure 7: Trace plots

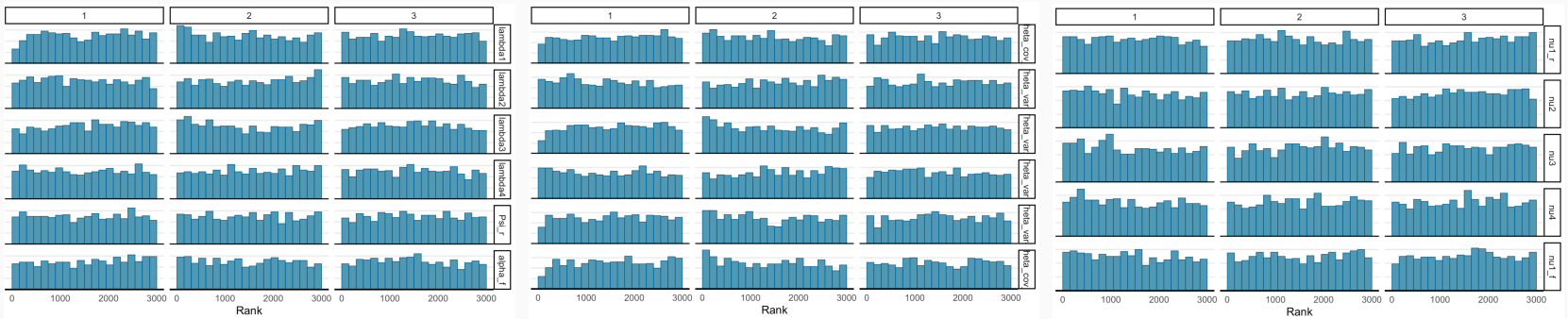


Figure 8: Rank histograms

Convergence: Neuroticism

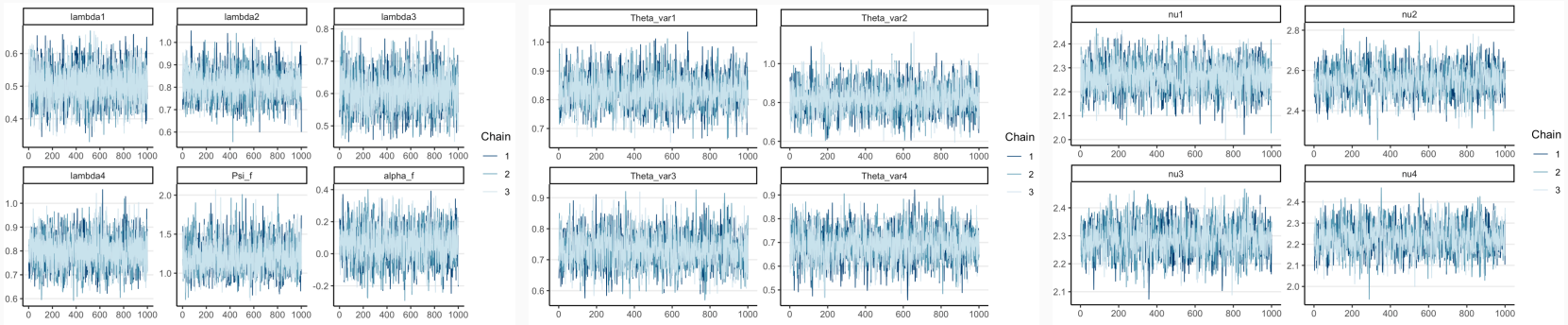


Figure 9: Trace plots

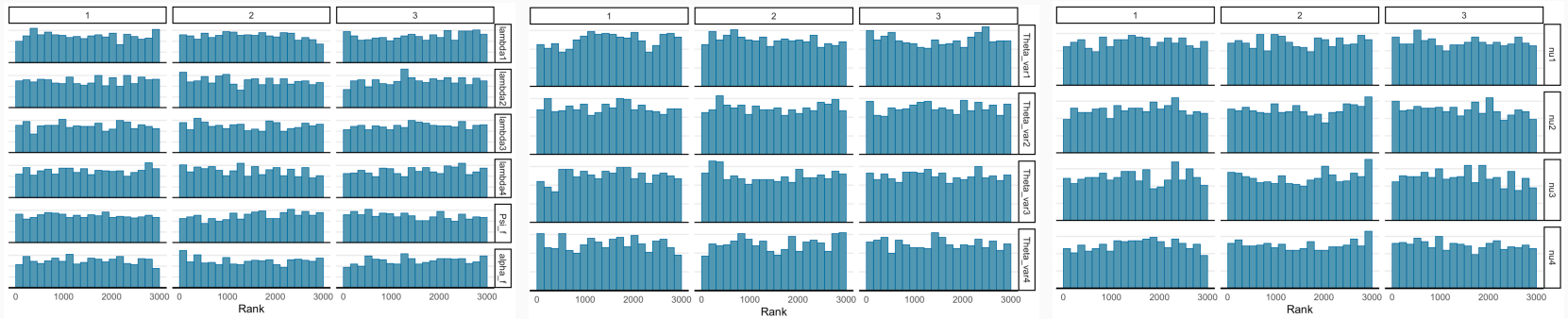


Figure 10: Rank histograms

Convergence: Openness to Experience

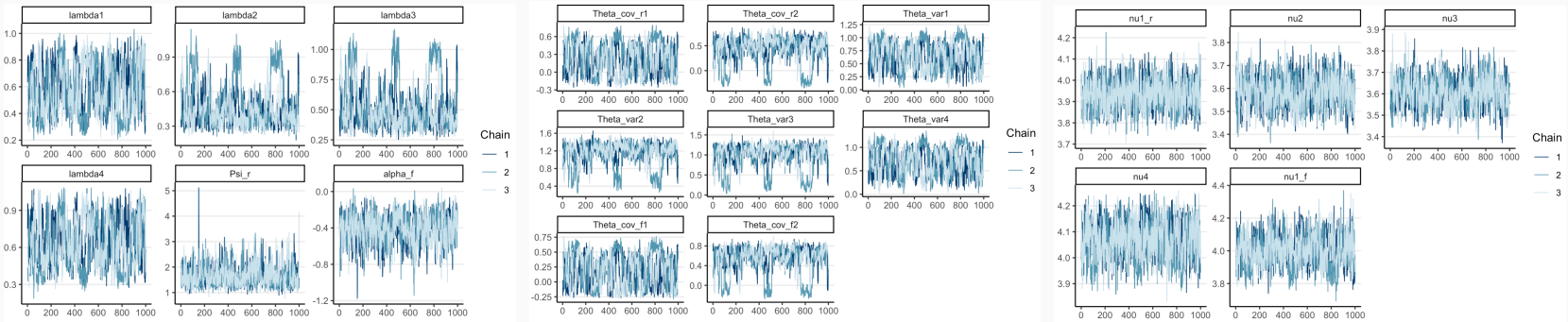


Figure 11: Trace plots

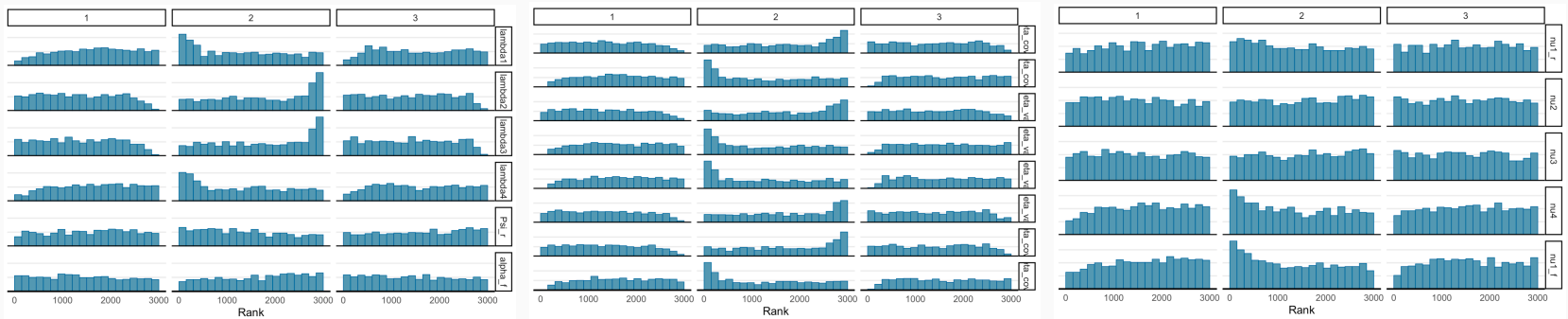


Figure 12: Rank histograms

Posterior summaries of classification accuracy indices

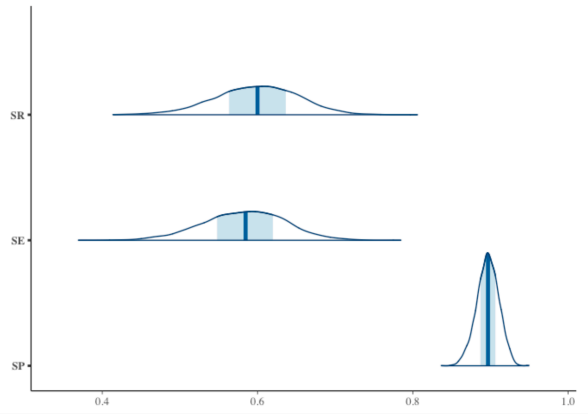
factor	variable	mean	median	sd	mad	q5	q95	rhat	ess_bulk	ess_tail
A	PS	0.15	0.15	0.00	0.00	0.15	0.15	1	3119	NA
	SR	0.57	0.58	0.05	0.05	0.49	0.65	1	2773	2852
	SE	0.57	0.58	0.05	0.05	0.49	0.65	1	2773	2852
	SP	0.92	0.92	0.01	0.01	0.91	0.94	1	2773	2852
C	PS	0.15	0.15	0.00	0.00	0.15	0.15	1	2874	NA
	SR	0.58	0.59	0.05	0.05	0.51	0.66	1	2423	2764
	SE	0.58	0.59	0.05	0.05	0.51	0.66	1	2423	2764
	SP	0.93	0.93	0.01	0.01	0.91	0.94	1	2423	2764
E	PS	0.15	0.15	0.00	0.00	0.15	0.15	1	2882	NA
	SR	0.66	0.66	0.04	0.03	0.60	0.73	1	2399	2849
	SE	0.66	0.66	0.04	0.03	0.60	0.73	1	2399	2849
	SP	0.94	0.94	0.01	0.01	0.93	0.95	1	2399	2849
N	PS	0.15	0.15	0.00	0.00	0.15	0.15	1	2595	NA
	SR	0.71	0.71	0.04	0.03	0.64	0.76	1	2526	1908
	SE	0.71	0.71	0.04	0.03	0.64	0.76	1	2526	1908
	SP	0.95	0.95	0.01	0.01	0.94	0.96	1	2526	1908
O	PS	0.15	0.15	0.00	0.00	0.15	0.15	1	2608	NA
	SR	0.51	0.52	0.06	0.07	0.40	0.61	1	427	780
	SE	0.51	0.52	0.06	0.07	0.40	0.61	1	427	780
	SP	0.91	0.91	0.01	0.01	0.89	0.93	1	427	780

Table 2: Posterior draw summaries for overall classification accuracy indices

CI for classification accuracy indices

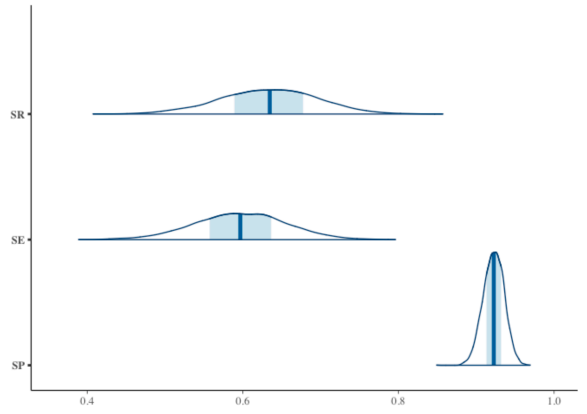
Agreeableness, group: F

```
mcmc_areas(agree_post_CAI$reference, pars = c( "SR", "SE", "SP"))
```



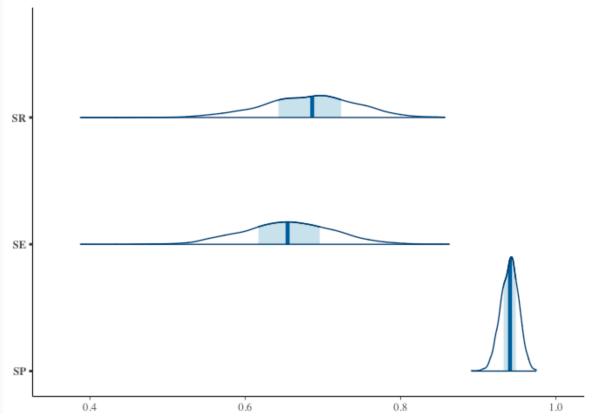
Conscientiousness, group: F

```
mcmc_areas(consc_post_CAI$reference, pars = c( "SR", "SE", "SP"))
```



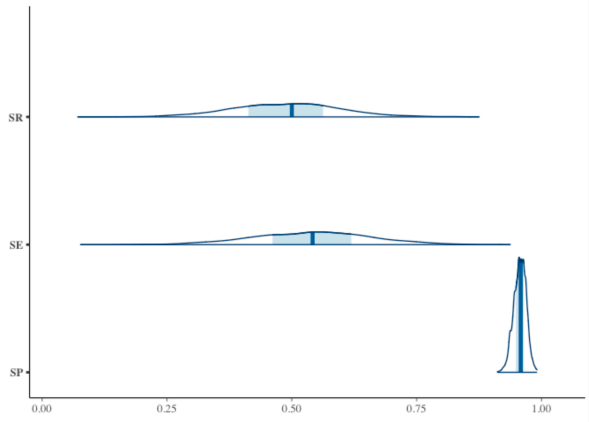
Extraversion, group: F

```
mcmc_areas(extr_post_CAI$reference, pars = c( "SR", "SE", "SP"))
```



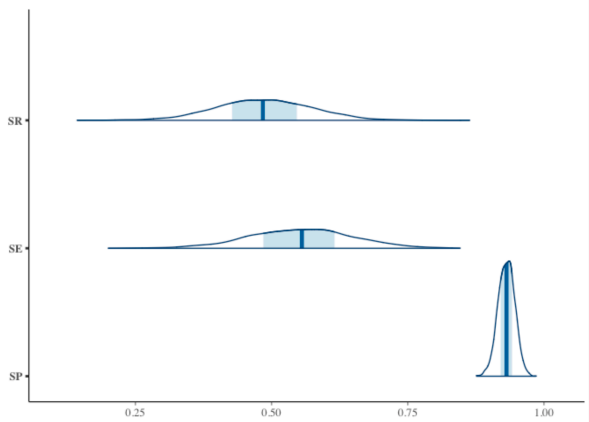
Agreeableness, group: M

```
mcmc_areas(agree_post_CAI$focal, pars = c( "SR", "SE", "SP"))
```



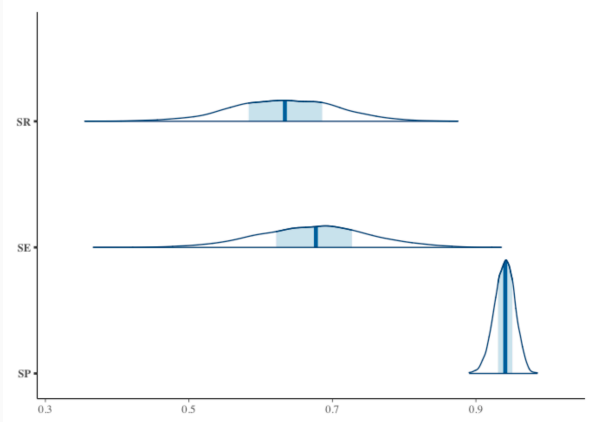
Conscientiousness, group: M

```
mcmc_areas(consc_post_CAI$focal, pars = c( "SR", "SE", "SP"))
```



Extraversion, group: M

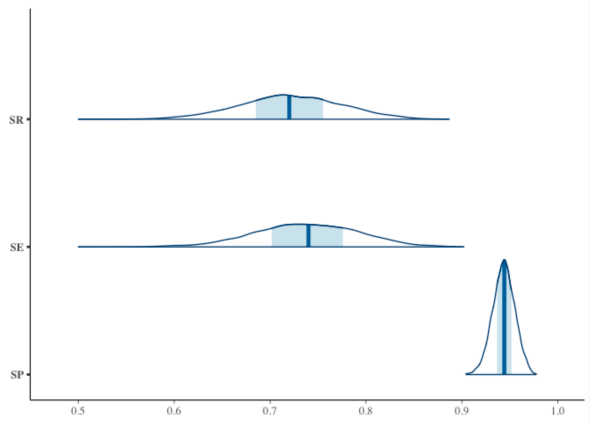
```
mcmc_areas(extr_post_CAI$focal, pars = c( "SR", "SE", "SP"))
```



CI for classification accuracy indices

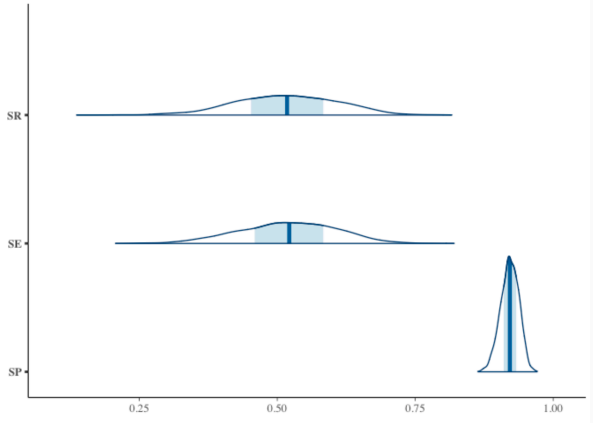
Neuroticism, group: F

```
mcmc_areas(neur_post_CAI$reference, pars = c( "SR", "SE", "SP"))
```



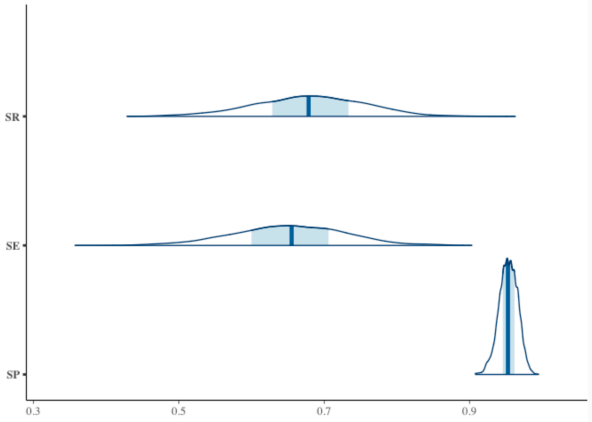
Openness to Experience, group: F

```
mcmc_areas(open_post_CAI$reference, pars = c( "SR", "SE", "SP"))
```



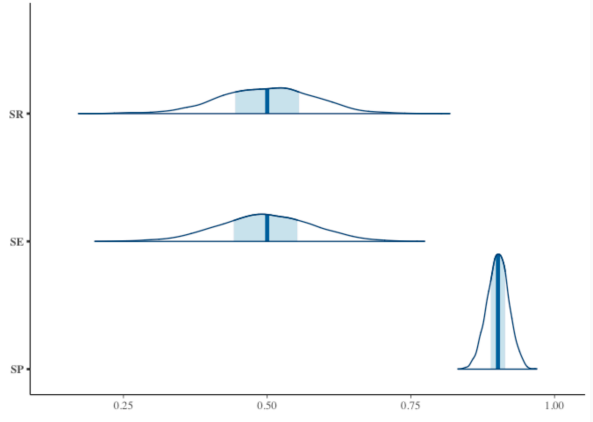
Neuroticism, group: M

```
mcmc_areas(neur_post_CAI$focal, pars = c( "SR", "SE", "SP"))
```



Openness to Experience, group: M

```
mcmc_areas(open_post_CAI$focal, pars = c( "SR", "SE", "SP"))
```



Posterior summaries for Cohen's h=2arcsin($\sqrt{p_1}$)-2arcsin($\sqrt{p_2}$)

factor	variable	mean	median	sd	mad	q5	q95	rhat	ess_bulk	ess_tail
A	h(PS)	0.37	0.36	0.08	0.07	0.23	0.51	1	2858	2816
	h(SR)	0.22	0.22	0.27	0.25	-0.21	0.65	1	2693	2768
	h(SE)	0.08	0.08	0.28	0.28	-0.39	0.55	1	3010	2742
	h(SP)	-0.24	-0.24	0.09	0.09	-0.39	-0.09	1	2714	2841
C	h(PS)	0.15	0.14	0.09	0.09	0.00	0.30	1	2406	2782
	h(SR)	0.30	0.30	0.25	0.25	-0.11	0.70	1	2631	2851
	h(SE)	0.09	0.08	0.25	0.25	-0.31	0.48	1	2428	2979
	h(SP)	-0.03	-0.03	0.09	0.09	-0.19	0.12	1	3169	3111
E	h(PS)	0.04	0.04	0.08	0.09	-0.08	0.19	1	1899	2349
	h(SR)	0.10	0.10	0.23	0.23	-0.27	0.47	1	2749	2980
	h(SE)	-0.04	-0.05	0.23	0.23	-0.42	0.33	1	2720	2899
	h(SP)	0.00	0.00	0.09	0.09	-0.15	0.15	1	2885	2740
N	h(PS)	0.12	0.12	0.07	0.06	0.00	0.25	1	2694	2776
	h(SR)	0.09	0.09	0.22	0.22	-0.27	0.44	1	3212	2742
	h(SE)	0.19	0.19	0.22	0.21	-0.18	0.56	1	2979	3110
	h(SP)	-0.05	-0.05	0.09	0.09	-0.20	0.10	1	2987	2784
O	h(PS)	-0.07	-0.06	0.09	0.09	-0.20	0.08	1	2259	2419
	h(SR)	0.03	0.04	0.25	0.25	-0.38	0.42	1	1090	1972
	h(SE)	0.05	0.04	0.22	0.23	-0.31	0.41	1	1972	2290
	h(SP)	0.07	0.07	0.10	0.09	-0.09	0.22	1	1698	2487

Table 3: Posterior draw summaries for Cohen's h for the difference in classification accuracy indices

CI for Cohen's h for the difference in classification accuracy indices

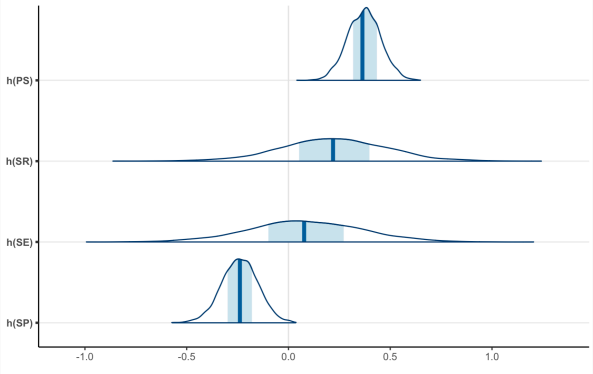


Figure 13: Agreeableness

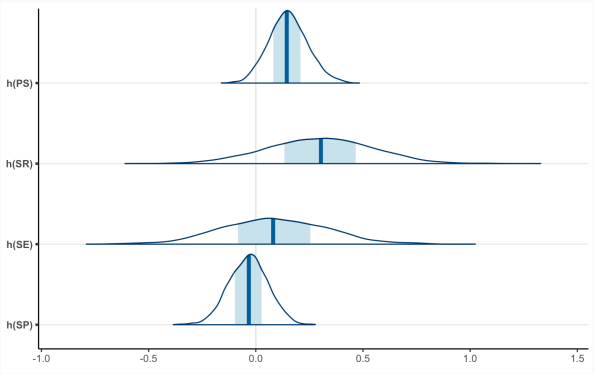


Figure 15: Conscientiousness

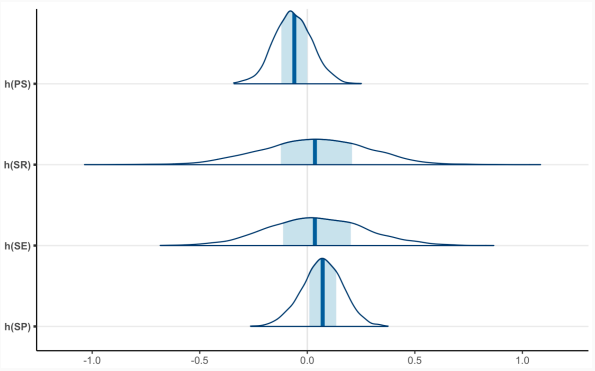


Figure 17: Openness to Experience

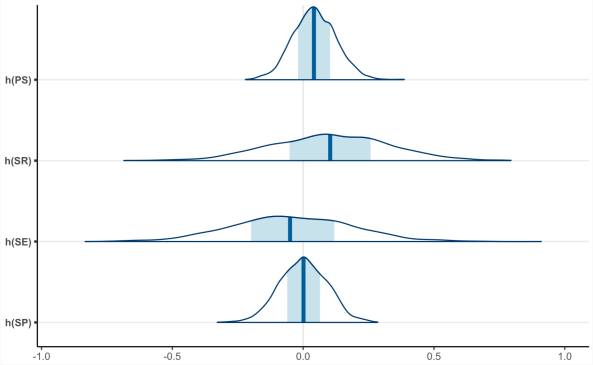


Figure 14: Extraversion

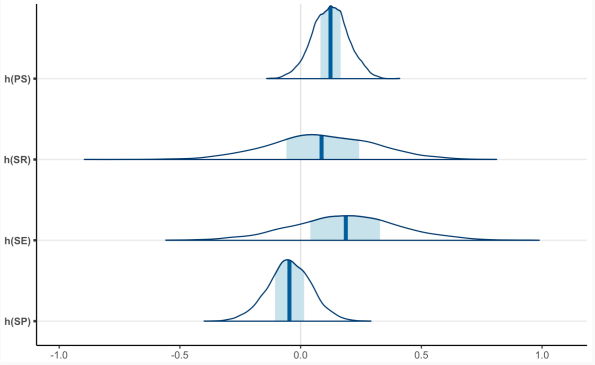


Figure 16: Neuroticism

Discussion

- Psychometric tests based on probabilistic inference -> error prone
- Measurement noninvariance may impact accuracy, cause adverse impact
- Wider credible intervals for SR and SE, greater confidence in PS and SP estimates
- Wider CI for h(SR) and h(SE), greater confidence in h(SP) estimate

Limitations and future directions:

- Only looked at subscales - selection often made with the full scale -> expand to multidimensional selection accuracy
- Compared the reference group to focal group - need to compare reference group to Efocal instead to better understand the impact of bias
- Openness to experience - different priors, longer chains, more thinning?
- Building Bayesian CI for the adverse impact (AI) ratio

Github

https://github.com/meltemozcan/573-Bayesian-final-project

References

- Byrne, B. M., Shavelson, R. J., & Muthén, B. (1989). Testing for the equivalence of factor covariance and mean structures: The issue of partial measurement invariance.. *Psychological Bulletin*, *105(3)*, 456–466. https://doi.org/10.1037/0033-2909.105.3.456
- Donnellan, M. B., Oswald, F. L., Baird, B. M., & Lucas, R. E. (2006). The mini-ipip scales: Tiny-yet-effective measures of the big five factors of personality. *Psychological Assessment*, *18*, 192–203. https://doi.org/doi:10.1037/1040-3590.18.2.192
- Horn, J. L., & McArdle, J. J. (1992). A practical and theoretical guide to measurement invariance in aging research.. *Experimental Aging Research*, *18:3*, 117–144. https://doi.org/10.1080/0361073920825391 6

Levy, R., & Mislevy, R. (2016). *Bayesian psychometric modeling (1st ed.)*. Chapman; Hall/CRC. https://doi.org/https://doi.org/10.1201/9781315374604

Mellenbergh, G. J. (1989). Item bias and item response theory. *International Journal of Educational Research*, *13*. https://doi.org/10.1016/0883-0355(89)90002-5

Meredith, W., & Teresi, J. A. (2006). An essay on measurement and factorial invariance. *Medical Care*, *44(11, Suppl 3)*. https://doi.org/10.1097/01.mlr.0000245438.73837.89

Merkle, E. C., Fitzsimmons, E., Uanhoro, J., & Goodrich, B. (2021). Efficient Bayesian structural equation modeling in Stan. *Journal of Statistical Software*, *100*(6), 1–22. https://doi.org/10.18637/jss.v100.i06

Millsap, R. E., & Kwok, O. M. (2004). Evaluating the impact of partial factorial invariance on selection in two populations. *Psychological Methods*, *9(1)*, 93–115. https://doi.org/https://doi.org/10.1037/1082-989X.9.1.93

Ock, J., McAbee, S. T., Mulfinger, E., & Oswald, F. L. (2020). The practical effects of measurement invariance: Gender invariance in two big five personality measures.. *Assessment*, *27(4)*, 657–674. https://doi.org/https://doi.org/10.1177/1073191119885018

Thurstone, L. L. (1947). *Multiple factor analysis*. University of Chicago Press.