IRT for teacher's practice

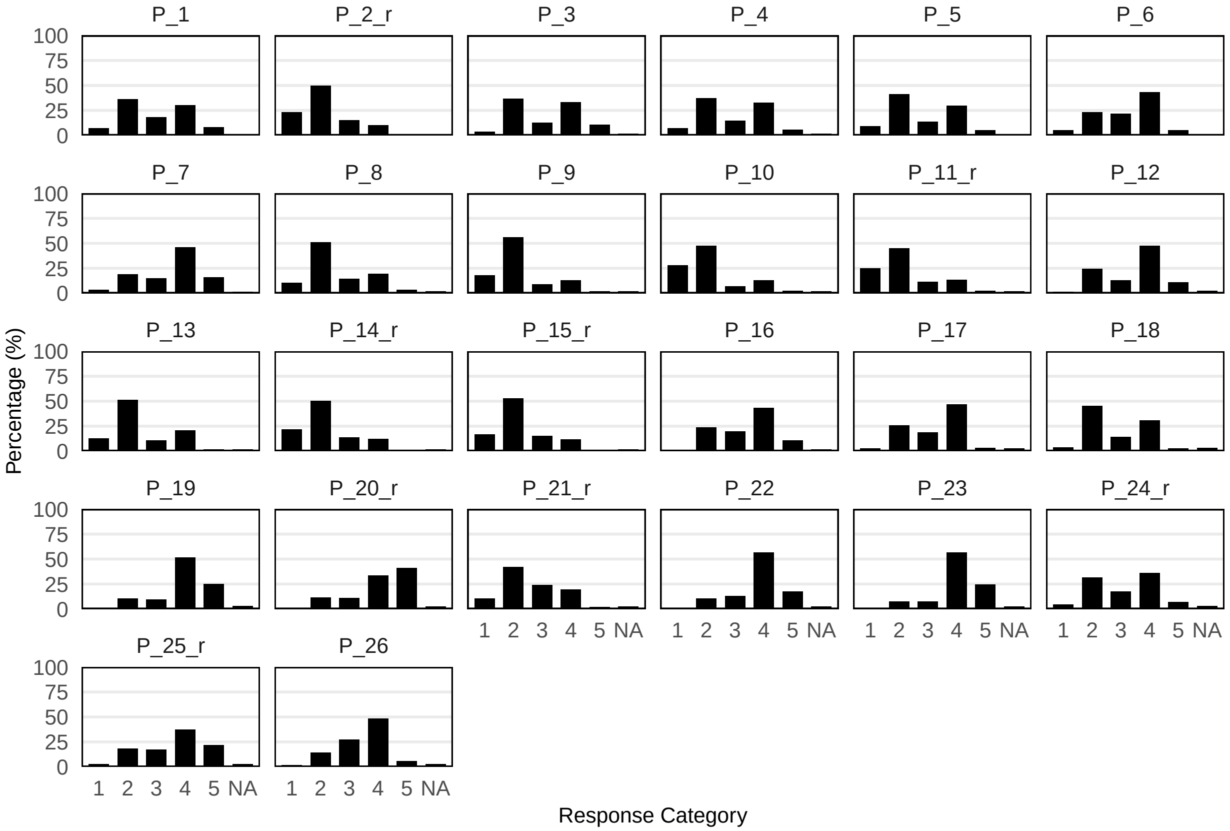
Chi Zhang

2025-01-30

The items and the responses

1	I introduce a new topic by first determining what the students already know about it.
2	I offer content matter in gradually increasing levels of complexity.
3	I jump between topics as the needs arise.
4	I have my students work collaboratively in pairs.
5	I have my students work collaboratively in groups.
6	I teach each student differently according to individual needs.
7	I encourage students to discuss the mistake they make.
8	Students work on extended mathematics investigations or projects.
9	Students work on projects in which subject material from various subjects is integrated.
10	Students make formal presentations to the rest of the class.
11	Students start with easy questions and work up to hard questions.
12	Students use mathematics concepts to interpret and solve applied problems.
13	Students play mathematical games.
14	Students work through exercises from textbooks or worksheets.
15	Students work on their own, consulting a neighbour from time to time.
16	I choose examples that appeal to students.
17	I try to indicate the value of each lesson topic for future use.
18	I encourage students to make connections to mathematical concepts that may be encountered in other areas of the curriculum.
19	When a student asks a question, I give a clue instead of the correct answer.
20	Students use only the methods I teach them.
21	During instruction I ask a lot of short questions to check whether students understand the content matter.
22	I ask students to explain their reasoning when giving an answer.
23	I encourage students to explore alternative methods for solution.
24	I avoid students making mistakes by explaining things carefully first.
25	I go through only one method for doing each question.
26	I allow students to work at their own pace.

This questionnniare is designed for assessing teacher's pedagogic practice, whether with connectionist/transmissionist tendency. Teachers were asked to response this on a 'frequency scale' manner, with 1-never, 2-seldom, 3-around half of the time, 4 - usually, 5 - almost always. There are items with reversed wording, so I first reversely code these items and see the whole response distribution.



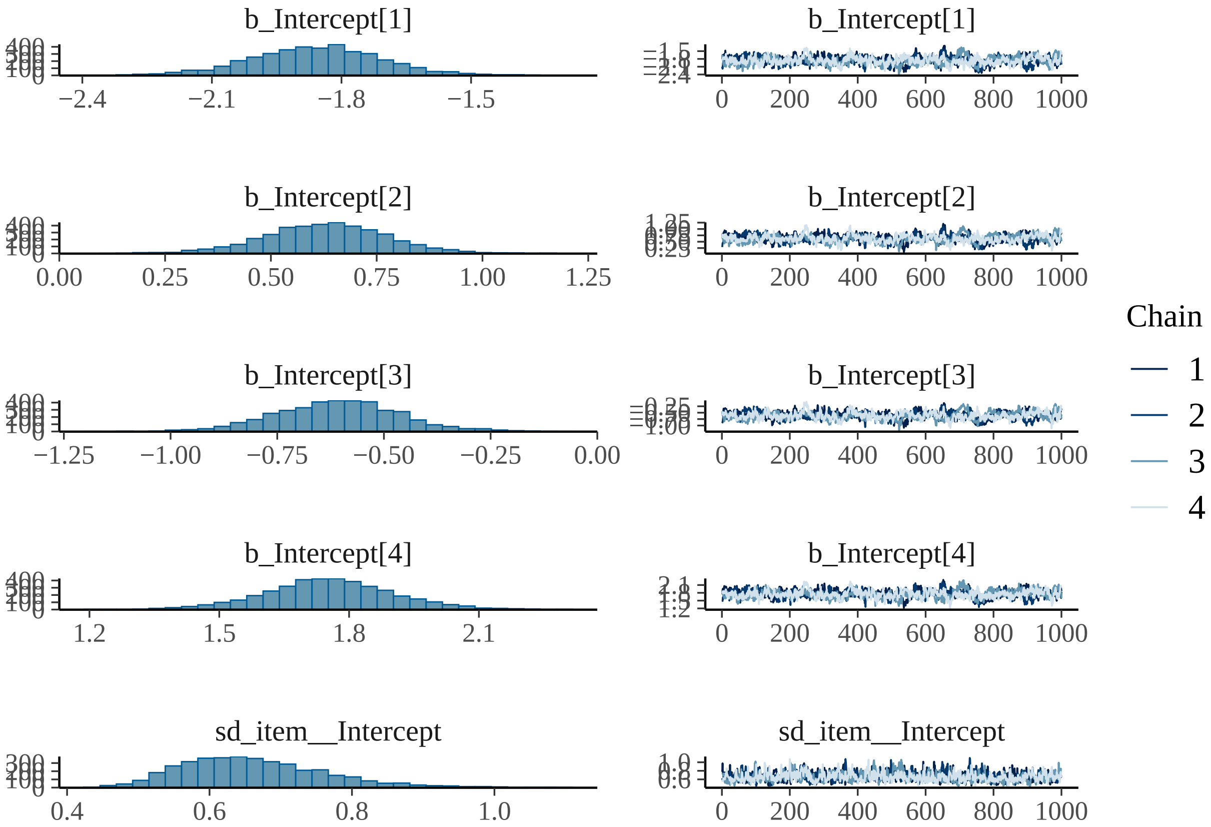
The distribution seems ‘OK’. It is not too bad as it has a reasonable distribution overall, but not too good as some items missing certain category (‘1’ or ‘5’). It is still acceptable considering our sample size is small. Another concern is that the category ‘3 - around half of the time’ seems not establish a ‘ordered’ distinction among others - many items with a ‘wave’ distribution where ‘2’ and ‘4’ are more likely to be endoresed than ‘3’. For now, we keep this original settings, but will come back to this issue later. Also there are three respondents miss more than 10 items. I’ll use the same weighted penalty strategy to treat this.

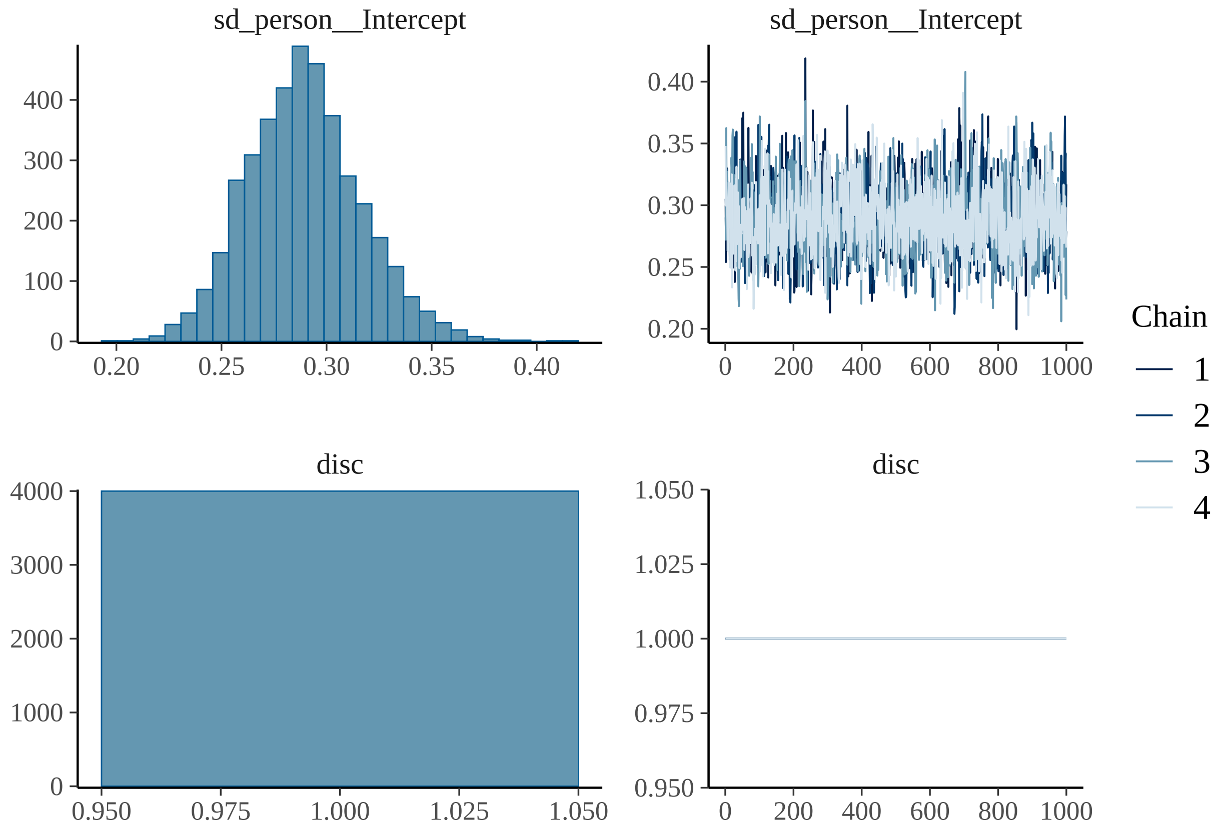
Foundational RSM model

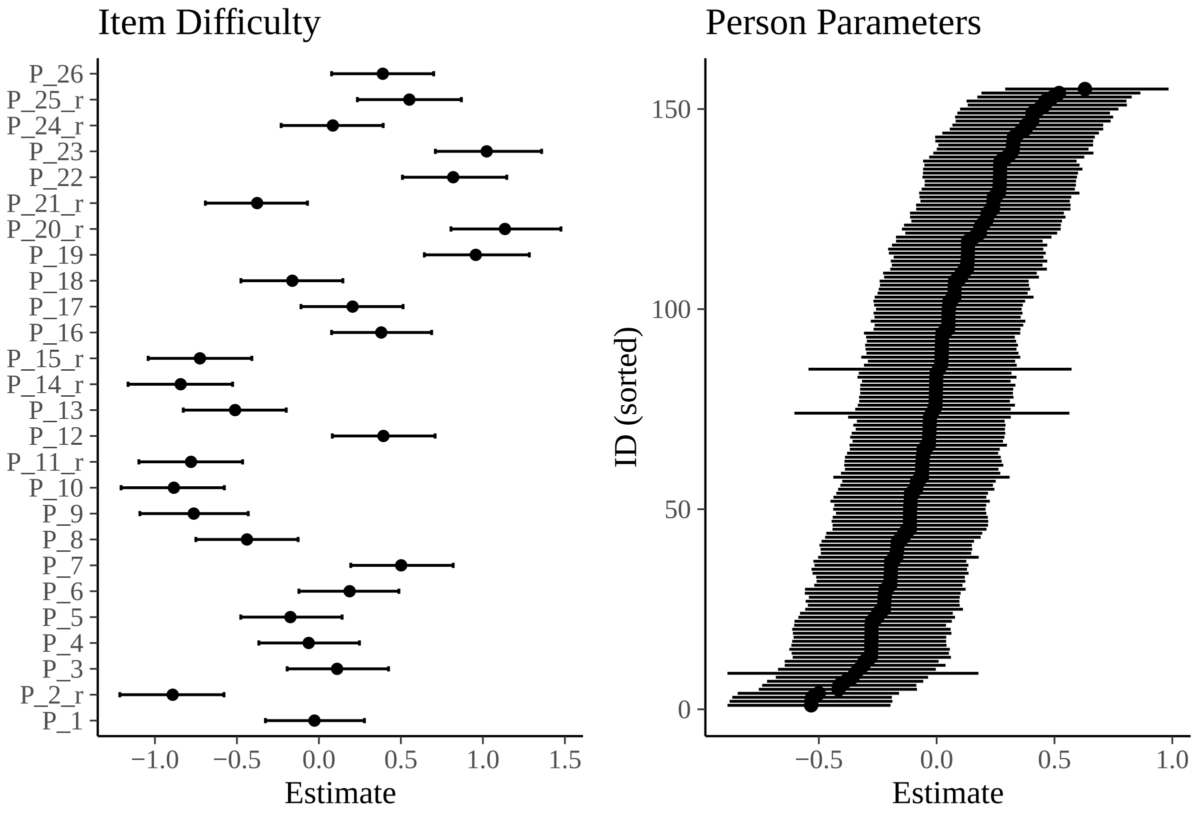
```
## [1] "Model Summary:"

##  Family: acat 
##   Links: mu = logit; disc = identity 
## Formula: response | weights(weight) ~ 1 + (1 | item) + (1 | person) 
##    Data: long_data_weighted (Number of observations: 3945) 
##   Draws: 4 chains, each with iter = 2000; warmup = 1000; thin = 1;
##          total post-warmup draws = 4000
## 
## Multilevel Hyperparameters:
## ~item (Number of levels: 26) 
##               Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
## sd(Intercept)     0.66      0.10     0.50     0.89 1.01      544     1166
## 
## ~person (Number of levels: 155) 
##               Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
## sd(Intercept)     0.29      0.03     0.24     0.35 1.00     1475     2301
##
## Regression Coefficients:
##              Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
## Intercept[1]    -1.86      0.15    -2.16    -1.56 1.01      259      529
## Intercept[2]     0.63      0.15     0.34     0.91 1.01      239      504
## Intercept[3]    -0.61      0.15    -0.89    -0.31 1.01      237      458
## Intercept[4]     1.75      0.15     1.46     2.06 1.01      258      529
##
## Further Distributional Parameters:
##      Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
## disc     1.00      0.00     1.00     1.00   NA       NA       NA
##
## Draws were sampled using sampling(NUTS). For each parameter, Bulk_ESS
## and Tail_ESS are effective sample size measures, and Rhat is the potential
## scale reduction factor on split chains (at convergence, Rhat = 1).
```

```
## [1] "\nModel Plot:"
```




sd_person__Intercept
400
300
200
100
0
0.20
0.25
0.30
0.35
0.40
sd_person__Intercept
0.40
0.35
0.30
0.25
0.20
0
200
400
600
800
1000
Chain
1
2
3
4
disc
4000
3000
2000
1000
0
0.950
0.975
1.000
1.025
1.050
disc
1.050
1.025
1.000
0.975
0.950
0
200
400
600
800
1000



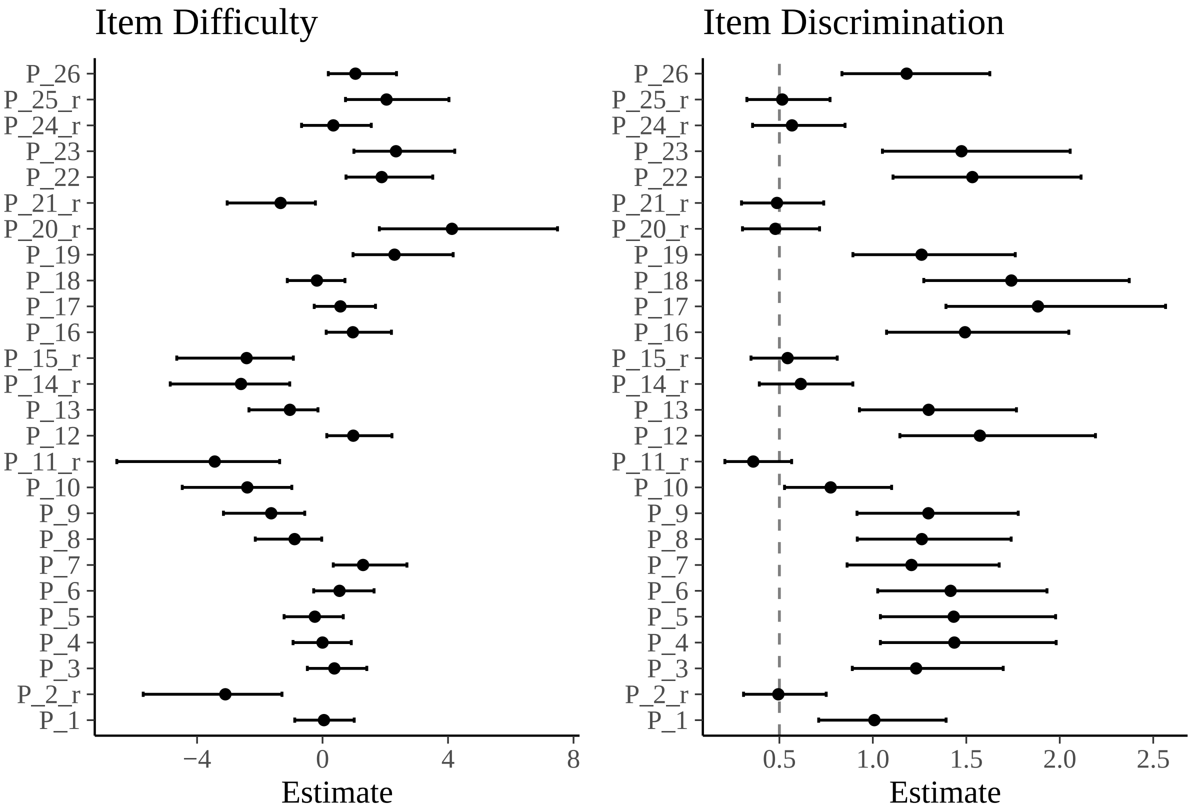
Starting from very basic 1PL setting - rsm, the difficulty/ability plot seems reasonable. But we can see that among several thresholds, 2 and 3 are reversed. There means that the difficulty/probability from transcending 2-3, is even harder than transcending from 3-4. This is problematic. Potential reasons can be teachers can't tell the ordered implications within these categogries - especially 3 is a category with explicit frequency referring (half of the time), while others are more like the vague feeling/perception of frequency. In any case, I would like to maintain the original category settings and see whether we can use differnet IRT structures to solve the problem. ## Introducing discrimination - GRSM model (2PL)

GRSM model summary

In the model summary, we can see that bringing discrimination parameter in even makes the gap between threshold 2&3 bigger. We might admit the threshold reverse is something inherint in the dataset, where partcipants failed to distinguish the order of the categories, or more likely, they failed to identify category 3 properly here.

```
##  Family: acat 
##   Links: mu = logit; disc = log 
## Formula: response ~ 1 + (1 | i | item) + (1 | person) 
##          disc ~ 1 + (1 | i | item)
##    Data: long_data_weighted (Number of observations: 3945) 
##   Draws: 4 chains, each with iter = 2000; warmup = 1000; thin = 1;
##          total post-warmup draws = 4000
## 
## Multilevel Hyperparameters:
## ~item (Number of levels: 26)
##                               Estimate Est.Error l-95% CI u-95% CI Rhat
## sd(Intercept)                     1.99      0.66     0.90     3.43 1.00
## sd(disc_Intercept)                0.55      0.10     0.39     0.76 1.00
## cor(Intercept,disc_Intercept)     0.33      0.20    -0.10     0.67 1.00
##                               Bulk_ESS Tail_ESS
## sd(Intercept)                      989     1615
## sd(disc_Intercept)                1334     1921
## cor(Intercept,disc_Intercept)     1296     2004
##
## ~person (Number of levels: 155)
##               Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
## sd(Intercept)     1.03      0.32     0.48     1.70 1.00     1070     1566
##
## Regression Coefficients:
##                Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
## Intercept[1]      -4.31      1.37    -7.29    -2.01 1.00     1015     1699
## Intercept[2]       1.27      0.58     0.35     2.59 1.01      561     1094
## Intercept[3]      -0.86      0.53    -2.05     0.03 1.01      546      885
## Intercept[4]       4.04      1.28     1.86     6.77 1.00      938     1562
## disc_Intercept    -0.82      0.33    -1.40    -0.11 1.00     1018     1438
##
## Draws were sampled using sampling(NUTS). For each parameter, Bulk_ESS
## and Tail_ESS are effective sample size measures, and Rhat is the potential
## scale reduction factor on split chains (at convergence, Rhat = 1).
```

GRSM difficulty/discrimination map



Some interesting results in discrimination plot also tell us the reversed items works unwell for our sample. take a look on them:

#	Item Description
2	I offer content matter in gradually increasing levels of complexity.
11	Students start with easy questions and work up to hard questions.
14	Students work through exercises from textbooks or worksheets.
15	Students work on their own, consulting a neighbour from time to time.
20	Students use only the methods I teach them.
21	During instruction I ask a lot of short questions to check whether students understand the content matter.
24	I avoid students making mistakes by explaining things carefully first.
25	I go through only one method for doing each question.

Basically, these items are used to assess teachers' transmissionist practice. Implicitly in this questionanire assumption, we assume that transmissionism and connectionnism are at the two end of a linear specturm, thus the more frequently teachers practice these, the more likely they are less connectionist. But here, it seems like all these transmissionist items have weak discrimination (taking 0.5 as a warning threshold) - that they can't function properly on distinguishing teachers' connectionnist tendency, that perhaps even though teachers with high connectionist score will still frequnetly condcut these transmissionist practices, and vice versa.

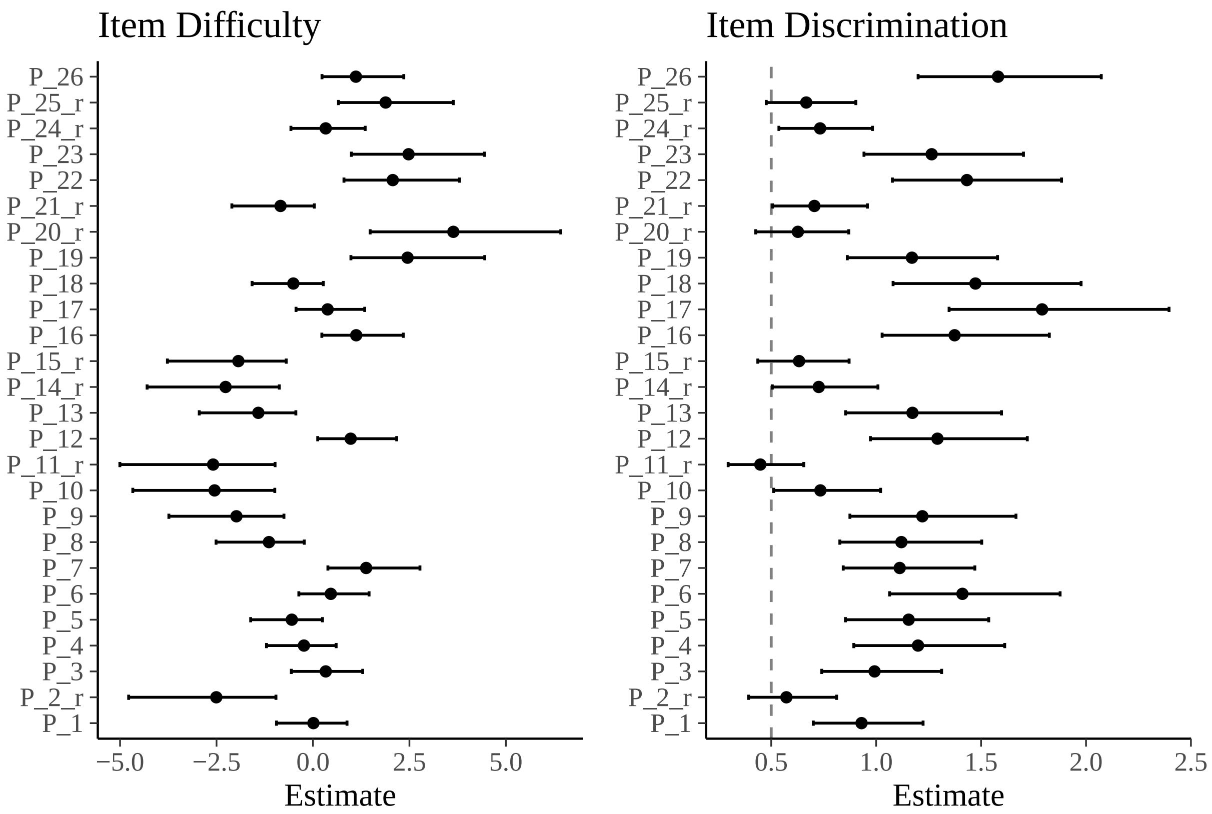
Model Iteration (category collapse)

Let's first solve the category reversing problem. I choose to collapse category 3 into 4, thus we have a new category as 'usually (around of more than half of the time)'. Technically it is also OK to integrate 2 and 3, but 'around half of the time' is sort of expression that closer to often/usually, linguistically people hardly regard 'half of the time' as 'seldom'.

GRSM_Merged stats

```
##  Family: acat
##   Links: mu = logit; disc = log
## Formula: response | weights(weight) ~ 1 + (1 | i | item) + (1 | person)
##          disc ~ 1 + (1 | i | item)
##    Data: long_data_merged (Number of observations: 3945)
##   Draws: 4 chains, each with iter = 2000; warmup = 1000; thin = 1;
##          total post-warmup draws = 4000
##
## Multilevel Hyperparameters:
## ~item (Number of levels: 26)
##                               Estimate Est.Error l-95% CI u-95% CI Rhat
## sd(Intercept)                     1.85      0.65     0.75     3.32 1.00
## sd(disc_Intercept)                0.41      0.08     0.29     0.59 1.00
## cor(Intercept,disc_Intercept)     0.31      0.20    -0.12     0.66 1.00
##                               Bulk_ESS Tail_ESS
## sd(Intercept)                      982     1294
## sd(disc_Intercept)                1325     2234
## cor(Intercept,disc_Intercept)     1410     2231
##
## ~person (Number of levels: 155)
##               Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
## sd(Intercept)     1.02      0.33     0.43     1.73 1.00     1116     1486
##
## Regression Coefficients:
##                Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
## Intercept[1]      -3.93      1.33    -6.86    -1.64 1.00      878     1356
## Intercept[2]      -0.59      0.45    -1.63     0.15 1.01      405      867
## Intercept[3]       4.03      1.33     1.71     6.89 1.00     1032     1439
## disc_Intercept    -0.52      0.35    -1.13     0.27 1.00     1092     1460
##
## Draws were sampled using sampling(NUTS). For each parameter, Bulk_ESS
## and Tail_ESS are effective sample size measures, and Rhat is the potential
## scale reduction factor on split chains (at convergence, Rhat = 1).
```

GRSM_Merged difficulty/discrimination map



After collapsing the categories, the output looks better - now the models has three ordered threshold with reasonable step-length, and the discrimination looks better. I'll try to maintain the collapsed setting and test the alternative IRT choices, i.e. GPCM (which provides flexible threshold within items), GRM (which with a different linking function forcing a ordered threshold setting), and PCM (if the item discrimination don't differ that much, PCM is a simpler version of GPCM and might works better for less parameters.)

Model Iteration (competing IRTs)

Model Comparasion (grsm vs grm vs gpcm vs pcm)

```
##  Family: acat
##   Links: mu = logit; disc = log
## Formula: response | weights(weight) ~ 1 + (cs(1) | i | item) + (1 | person)
##          disc ~ 1 + (1 | i | item)
##    Data: long_data_merged (Number of observations: 3945)
##   Draws: 4 chains, each with iter = 2000; warmup = 1000; thin = 1;
##          total post-warmup draws = 4000
##
## Multilevel Hyperparameters:
## ~item (Number of levels: 26)
##                                   Estimate Est.Error l-95% CI u-95% CI Rhat
## sd(Intercept[1])                     2.05      0.72     0.92     3.72 1.00
## sd(Intercept[2])                     2.20      0.70     1.04     3.77 1.00
## sd(Intercept[3])                     4.24      1.32     2.05     7.20 1.00
## sd(disc_Intercept)                   0.87      0.13     0.65     1.17 1.00
## cor(Intercept[1],Intercept[2])       0.59      0.18     0.19     0.86 1.01
## cor(Intercept[1],Intercept[3])      -0.13      0.21    -0.51     0.30 1.00
## cor(Intercept[2],Intercept[3])       0.68      0.13     0.37     0.87 1.00
## cor(Intercept[1],disc_Intercept)    -0.65      0.13    -0.86    -0.37 1.00
## cor(Intercept[2],disc_Intercept)     0.14      0.21    -0.28     0.53 1.00
## cor(Intercept[3],disc_Intercept)     0.75      0.09     0.53     0.88 1.00
##                                  Bulk_ESS Tail_ESS
## sd(Intercept[1])                      793     1383
## sd(Intercept[2])                      862     1535
## sd(Intercept[3])                      880     1324
## sd(disc_Intercept)                    832     1749
## cor(Intercept[1],Intercept[2])        567     1124
## cor(Intercept[1],Intercept[3])        616     1037
## cor(Intercept[2],Intercept[3])       1332     2546
## cor(Intercept[1],disc_Intercept)      824     1896
## cor(Intercept[2],disc_Intercept)     1002     2160
## cor(Intercept[3],disc_Intercept)     1163     2103
##
## ~person (Number of levels: 155)
##               Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
## sd(Intercept)     0.74      0.23     0.35     1.26 1.00      702     1228
##
## Regression Coefficients:
##                Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
## Intercept[1]      -3.35      1.04    -5.62    -1.62 1.00      744     1225
## Intercept[2]      -0.47      0.50    -1.52     0.38 1.00      548      959
## Intercept[3]       4.10      1.27     1.92     6.86 1.00     1037     1505
## disc_Intercept    -0.21      0.32    -0.79     0.49 1.00      856     1543
##
## Draws were sampled using sampling(NUTS). For each parameter, Bulk_ESS
## and Tail_ESS are effective sample size measures, and Rhat is the potential
## scale reduction factor on split chains (at convergence, Rhat = 1).
```

```
##  Family: cumulative
##   Links: mu = logit; disc = log
## Formula: response | weights(weight) ~ 1 + (1 | i | item) + (1 | person)
##          disc ~ 1 + (1 | i | item)
##    Data: long_data_merged (Number of observations: 3945)
##   Draws: 4 chains, each with iter = 2000; warmup = 1000; thin = 1;
##          total post-warmup draws = 4000
##
## Multilevel Hyperparameters:
## ~item (Number of levels: 26)
##                              Estimate Est.Error l-95% CI u-95% CI Rhat
## sd(Intercept)                    1.76      0.61     0.76     3.14 1.00
## sd(disc_Intercept)               0.23      0.04     0.16     0.32 1.00
## cor(Intercept,disc_Intercept)    0.31      0.20    -0.11     0.66 1.00
##                              Bulk_ESS Tail_ESS
## sd(Intercept)                     865      859
## sd(disc_Intercept)               1343     2321
## cor(Intercept,disc_Intercept)    1325     1721
##
```

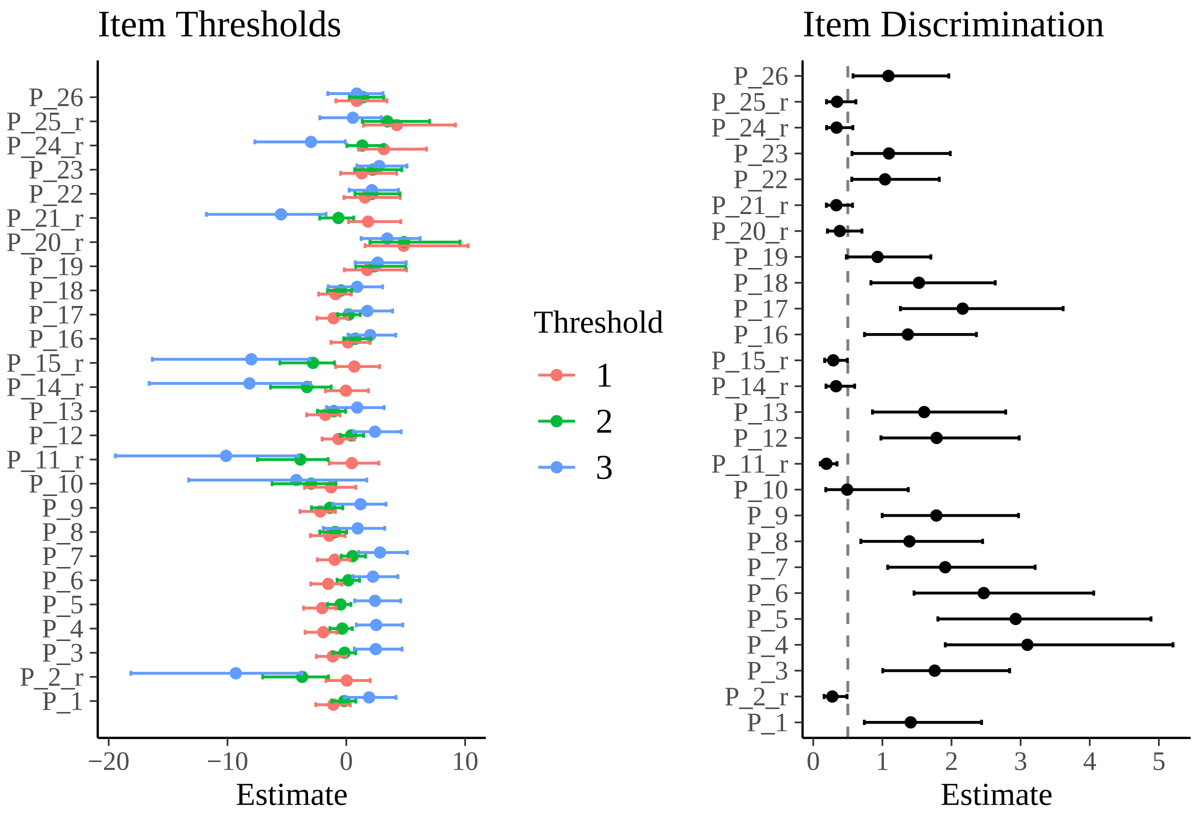
```
## ~person (Number of levels: 155)
##               Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
## sd(Intercept)     0.97      0.32     0.44     1.67 1.00      915     1152
##
## Regression Coefficients:
##                Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
## Intercept[1]      -4.20      1.38    -7.22    -1.87 1.00      894     1052
## Intercept[2]      -0.47      0.39    -1.34     0.21 1.00      509      831
## Intercept[3]       3.96      1.32     1.79     6.90 1.00      847     1033
## disc_Intercept    -0.25      0.34    -0.83     0.48 1.00      911     1020
##
## Draws were sampled using sampling(NUTS). For each parameter, Bulk_ESS
## and Tail_ESS are effective sample size measures, and Rhat is the potential
## scale reduction factor on split chains (at convergence, Rhat = 1).
```

```
##  Family: acat
##   Links: mu = logit; disc = log
## Formula: response | weights(weight) ~ 1 + (cs(1) | i | item) + (1 | person)
##          disc ~ 1 + (1 | i | item)
##    Data: long_data_merged (Number of observations: 3945)
##   Draws: 4 chains, each with iter = 2000; warmup = 1000; thin = 1;
##          total post-warmup draws = 4000
##
## Multilevel Hyperparameters:
## ~item (Number of levels: 26)
##                                  Estimate Est.Error l-95% CI u-95% CI Rhat
## sd(Intercept[1])                     2.06      0.75     0.93     3.78 1.00
## sd(Intercept[2])                     2.21      0.73     1.03     3.88 1.00
## sd(Intercept[3])                     4.30      1.38     2.05     7.32 1.00
## sd(disc_Intercept)                   0.88      0.13     0.65     1.17 1.00
## cor(Intercept[1],Intercept[2])       0.57      0.19     0.13     0.85 1.00
## cor(Intercept[1],Intercept[3])      -0.14      0.21    -0.54     0.27 1.00
## cor(Intercept[2],Intercept[3])       0.68      0.12     0.40     0.87 1.00
## cor(Intercept[1],disc_Intercept)    -0.66      0.12    -0.86    -0.39 1.00
## cor(Intercept[2],disc_Intercept)     0.15      0.21    -0.27     0.54 1.00
## cor(Intercept[3],disc_Intercept)     0.75      0.09     0.56     0.88 1.00
##                                  Bulk_ESS Tail_ESS
## sd(Intercept[1])                     1301     1920
## sd(Intercept[2])                     1300     2057
## sd(Intercept[3])                     1373     2019
## sd(disc_Intercept)                   1089     2086
## cor(Intercept[1],Intercept[2])        688     1551
## cor(Intercept[1],Intercept[3])        638     1229
## cor(Intercept[2],Intercept[3])       1903     2916
## cor(Intercept[1],disc_Intercept)      982     1809
## cor(Intercept[2],disc_Intercept)     1493     2811
## cor(Intercept[3],disc_Intercept)     1417     2538
##
## ~person (Number of levels: 155)
##               Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
## sd(Intercept)     0.74      0.25     0.35     1.34 1.00     1037     1783
##
## Regression Coefficients:
##                Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
## Intercept[1]      -3.37      1.12    -5.96    -1.57 1.00     1095     1937
## Intercept[2]      -0.42      0.49    -1.54     0.43 1.01      791     1411
## Intercept[3]       4.23      1.48     1.83     7.57 1.00     1123     1476
## disc_Intercept    -0.22      0.35    -0.87     0.50 1.00     1068     1345
##
## Draws were sampled using sampling(NUTS). For each parameter, Bulk_ESS
## and Tail_ESS are effective sample size measures, and Rhat is the potential
## scale reduction factor on split chains (at convergence, Rhat = 1).
```

```
##      elpd_diff se_diff
## PCM     0.0       0.0
## GPCM   -0.3       0.4
## GRSM -123.4      12.8
## GRM  -133.2      14.2
```

All four models offer reasonable threshold and other stats. And the loo test tolds us that PCM & GPCM work similarly well, and both way better than GRM and GRSM.

GPCM check



While looking closely into GPCM, the output looks not that good anymore. We can see the flexible threshold setting let the model be more sensitive about abnormal data pattern - originally, grsm assume every item with identical threshold, while in GPCM map shows that the many items strongly against the assumption. More specifically, almost all reversed coding items shows strongly reverse threshold order, and some items have extremely unstable yet low estimate on threshold 3, implying for teachers with high connectionnist tendency, it is much easier to endorse category 3 (that frequently conducting transmissionist pedagogy) than category 1 (seldom conduct transmissionist practice). Moreover, the discrimination of these items are either lower or very likely lower than the warning threshold 0.5, suggesting that these item can't distinguish teachers well in terms of connectionnism (if we now can admit that the connectionnism is still on the spectrum while transmissionism might not at the linear other end any more.)

Model Iteration (adjusting items)

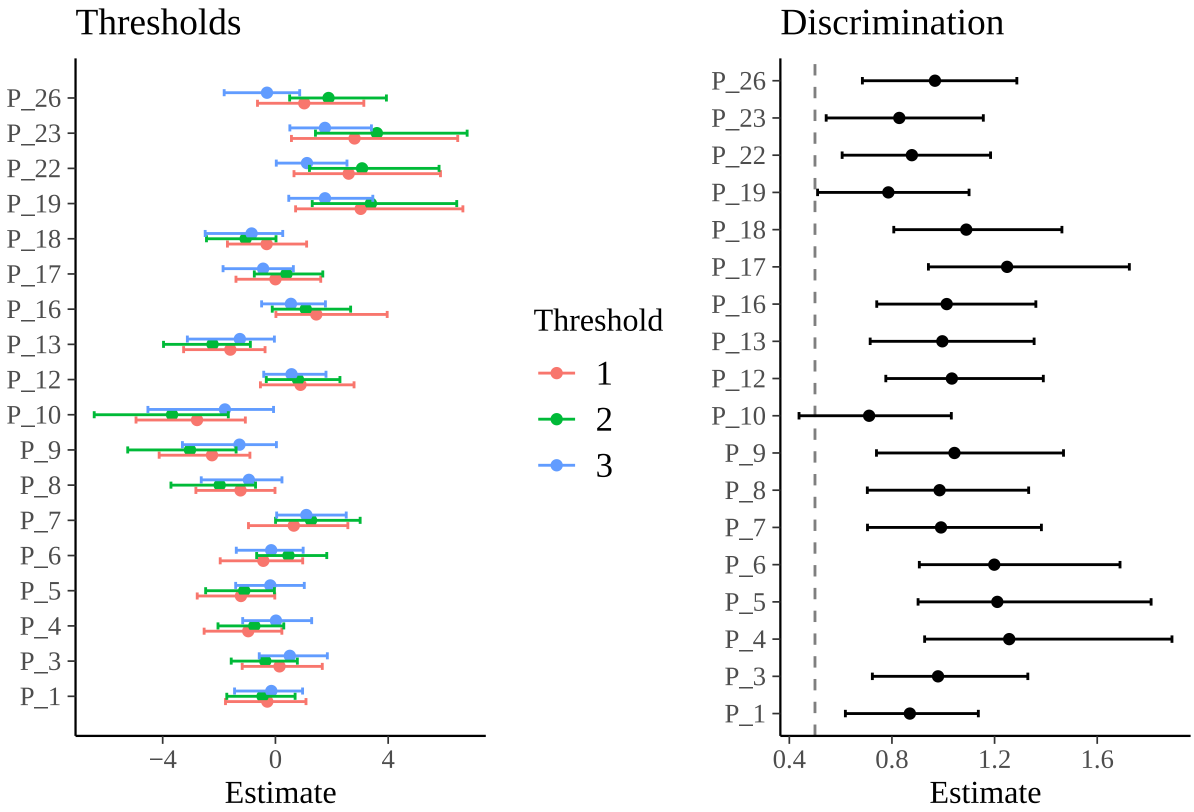
Stats of GPCM with filtered items

```
##  Family: acat
##   Links: mu = logit; disc = log
## Formula: response ~ 1 + (cs(1) | i | item) + (1 | person)
##          disc ~ 1 + (1 | i | item)
##    Data: long_data_filtered (Number of observations: 2733)
##   Draws: 4 chains, each with iter = 2000; warmup = 1000; thin = 1;
##          total post-warmup draws = 4000
##
## Multilevel Hyperparameters:
## ~item (Number of levels: 18)
##                                  Estimate Est.Error l-95% CI u-95% CI Rhat
## sd(Intercept[1])                     1.74      0.65     0.72     3.25 1.00
## sd(Intercept[2])                     2.09      0.67     0.96     3.60 1.00
## sd(Intercept[3])                     1.17      0.46     0.47     2.22 1.00
## sd(disc_Intercept)                   0.24      0.10     0.06     0.44 1.00
## cor(Intercept[1],Intercept[2])       0.85      0.12     0.52     0.98 1.00
## cor(Intercept[1],Intercept[3])       0.60      0.22     0.11     0.93 1.00
## cor(Intercept[2],Intercept[3])       0.66      0.19     0.21     0.93 1.00
## cor(Intercept[1],disc_Intercept)    -0.21      0.32    -0.72     0.53 1.00
## cor(Intercept[2],disc_Intercept)    -0.09      0.33    -0.69     0.57 1.00
## cor(Intercept[3],disc_Intercept)     0.03      0.36    -0.70     0.65 1.00
##                                  Bulk_ESS Tail_ESS
## sd(Intercept[1])                     1186     1545
## sd(Intercept[2])                     1086     1211
## sd(Intercept[3])                     1339     1902
## sd(disc_Intercept)                    772      819
## cor(Intercept[1],Intercept[2])       1610     2055
## cor(Intercept[1],Intercept[3])       1940     2716
## cor(Intercept[2],Intercept[3])       3077     3067
## cor(Intercept[1],disc_Intercept)     1579     2223
## cor(Intercept[2],disc_Intercept)     2130     3002
## cor(Intercept[3],disc_Intercept)     1619     2539
##
## ~person (Number of levels: 155)
##               Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
## sd(Intercept)     1.69      0.47     0.85     2.70 1.00     1108     1689
##
## Regression Coefficients:
##                Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
## Intercept[1]      -4.69      1.32    -7.51    -2.40 1.00     1161     1527
## Intercept[2]      -1.07      0.59    -2.35    -0.06 1.00     1407     2324
## Intercept[3]       4.22      1.21     2.05     6.88 1.00     1035     1651
## disc_Intercept    -0.42      0.29    -0.93     0.22 1.00     1142     1377
```

```
##
## Draws were sampled using sampling(NUTS). For each parameter, Bulk_ESS
## and Tail_ESS are effective sample size measures, and Rhat is the potential
## scale reduction factor on split chains (at convergence, Rhat = 1).
```

Now we delete all the reversed coding items and re-fit the GPCM model (I iterate for varies version to try to maintain some reversed item, but can't find a way to maintain all items certainly above the warning discrimination threshold 0.5.)

##GPCM_filtered difficulty-discrimination map



After adjustment, the data seems much more reasonable: firstly, no item's discrimination is completely below the warning threshold; secondly, the difficulty threshold of each item is more evenly distributed - although there are still some items that seem to have a reversed pattern within the threshold, there are no longer extreme outliers. Especially considering the small sample size, this iterative version is acceptable.

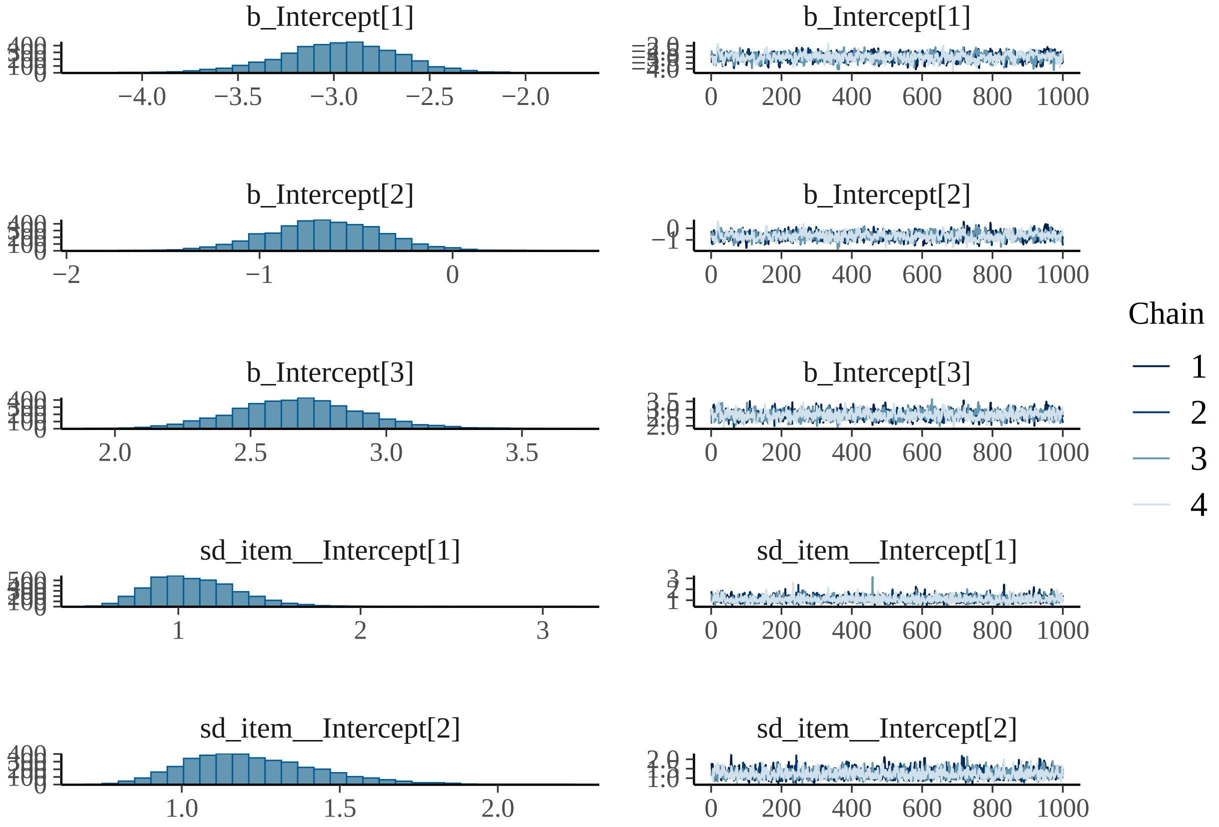
Competing IRTs

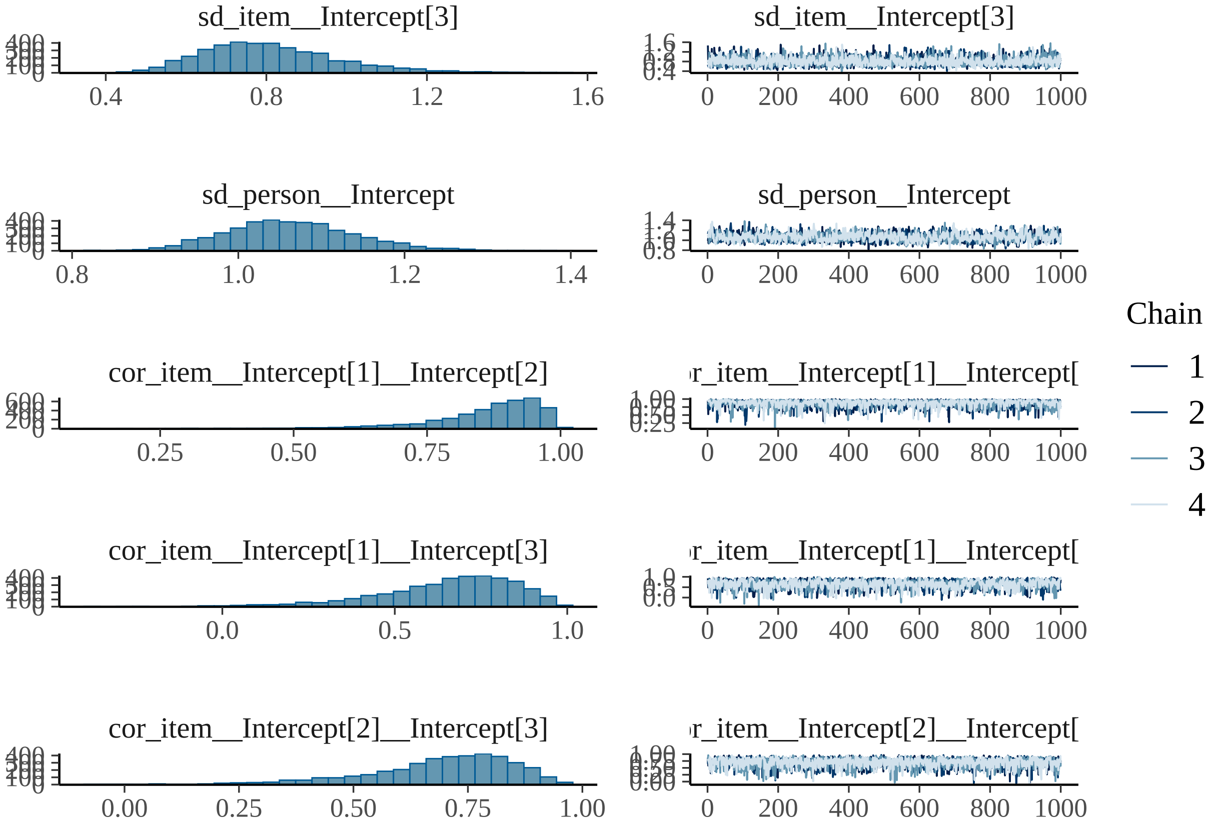
```
##       elpd_diff se_diff
## GRM    0.0       0.0
## GPCM -0.7       7.4
## PCM  -4.5       6.8
## GRSM -9.0       4.1
```

We can see now there is only very small difference between models - we can still say that GRSM works worse than others (|elpd_diff| > 2* se_diff), but basically GRM/GPCM/PCM works the same. From GPCM discrimination map, we already see that we have a relatively stable discrimination distribution, and from the loo test, we know that the model won't lose significant information by assuming all discriminations = 1. Thus I choose PCM model for the final structure.

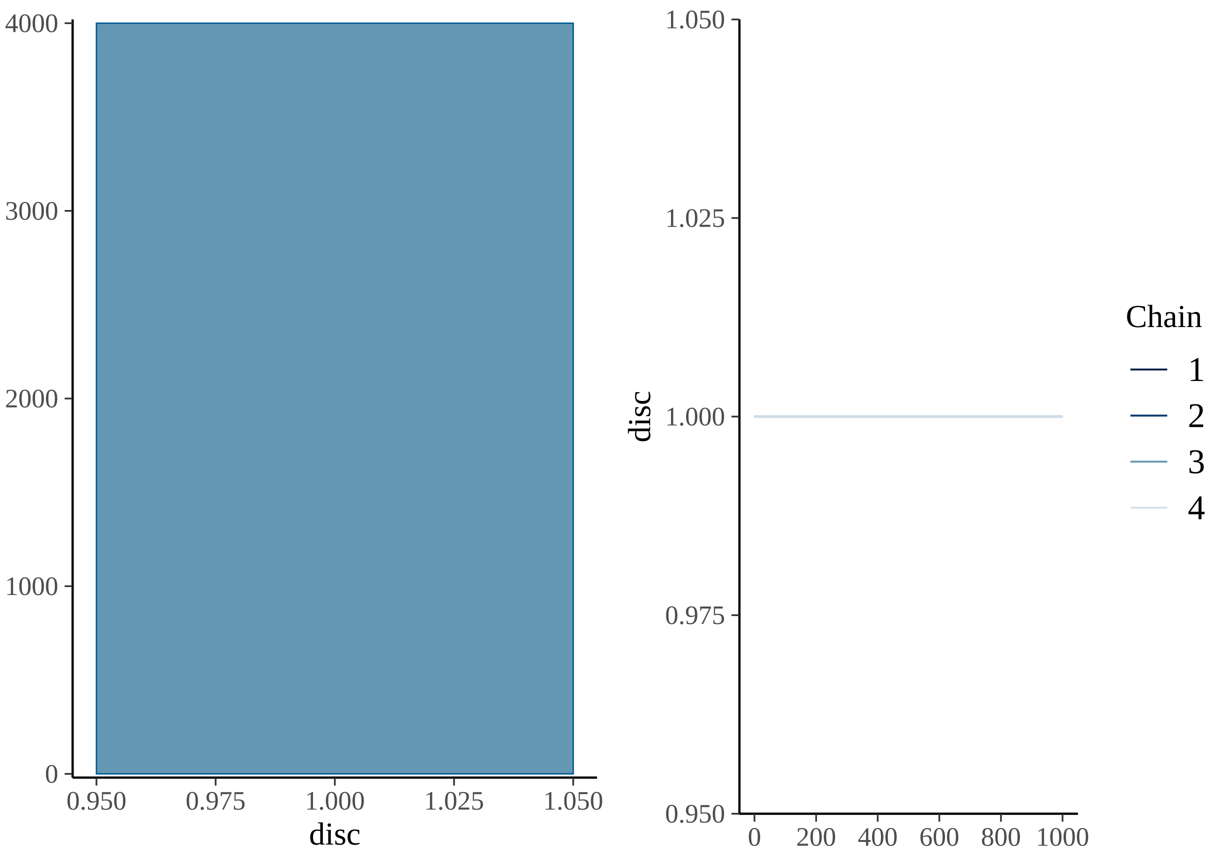
Final choice - PCM

```
##  Family: acat 
##   Links: mu = logit; disc = identity 
## Formula: response ~ 1 + (cs(1) | i | item) + (1 | person) 
##    Data: long_data_filtered (Number of observations: 2733) 
##   Draws: 4 chains, each with iter = 2000; warmup = 1000; thin = 1;
##          total post-warmup draws = 4000
## 
## Multilevel Hyperparameters:
## ~item (Number of levels: 18) 
##                               Estimate Est.Error l-95% CI u-95% CI Rhat
## sd(Intercept[1])                  1.09      0.25     0.69     1.64 1.00
## sd(Intercept[2])                  1.23      0.22     0.88     1.71 1.00
## sd(Intercept[3])                  0.81      0.17     0.54     1.21 1.00
## cor(Intercept[1],Intercept[2])    0.86      0.11     0.58     0.99 1.00
## cor(Intercept[1],Intercept[3])    0.65      0.20     0.17     0.93 1.00
## cor(Intercept[2],Intercept[3])    0.69      0.16     0.29     0.92 1.00
##                               Bulk_ESS Tail_ESS
## sd(Intercept[1])                  2051     2780
## sd(Intercept[2])                  1803     2194
## sd(Intercept[3])                  2392     2694
## cor(Intercept[1],Intercept[2])    1802     2642
## cor(Intercept[1],Intercept[3])    1400     2090
## cor(Intercept[2],Intercept[3])    2017     3084
## 
## ~person (Number of levels: 155) 
##               Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
## sd(Intercept)     1.06      0.08     0.92     1.23 1.00     1315     2364
## 
## Regression Coefficients:
##              Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
## Intercept[1]    -2.98      0.31    -3.64    -2.40 1.00     1100     1763
## Intercept[2]    -0.66      0.30    -1.25    -0.06 1.00     1059     1517
## Intercept[3]     2.68      0.24     2.22     3.17 1.00     1398     2153
## 
## Further Distributional Parameters:
##      Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
## disc     1.00      0.00     1.00     1.00   NA       NA       NA
## 
## Draws were sampled using sampling(NUTS). For each parameter, Bulk_ESS
## and Tail_ESS are effective sample size measures, and Rhat is the potential
## scale reduction factor on split chains (at convergence, Rhat = 1).
```

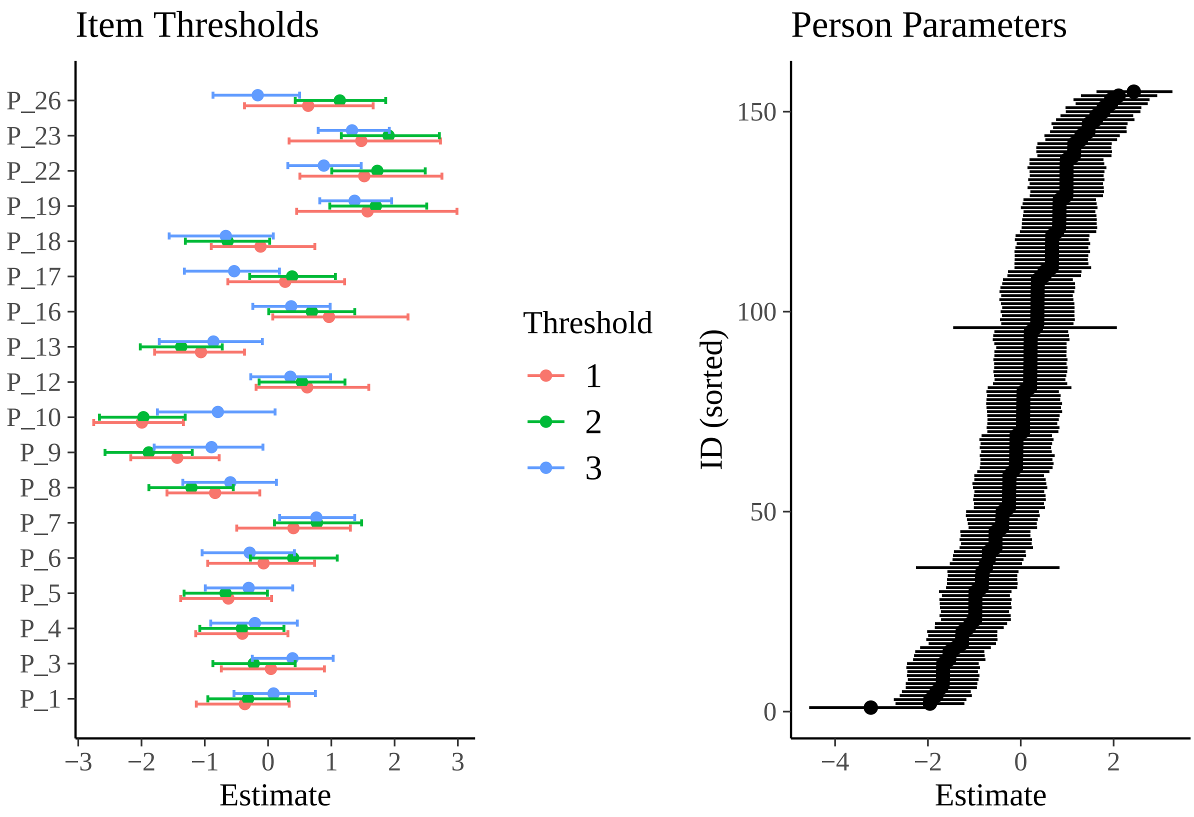

b_Intercept[1]
b_Intercept[1]
−4.0
−3.5
−3.0
−2.5
−2.0
0
200
400
600
800
1000
b_Intercept[2]
b_Intercept[2]
−2
−1
0
0
−1
b_Intercept[3]
b_Intercept[3]
2.0
2.5
3.0
3.5
sd_item__Intercept[1]
sd_item__Intercept[1]
1
2
3
sd_item__Intercept[2]
sd_item__Intercept[2]
1.0
1.5
2.0
Chain
1
2
3
4



sd_item__Intercept[3]
400
0
0.4
0.8
1.2
1.6
sd_item__Intercept[3]
1.6
0.4
0
200
400
600
800
1000
sd_person__Intercept
400
0
0.8
1.0
1.2
1.4
sd_person__Intercept
1.4
0.8
0
200
400
600
800
1000
Chain
1
2
3
4
cor_item__Intercept[1]__Intercept[2]
600
0
0.25
0.50
0.75
1.00
r_item__Intercept[1]__Intercept[
1.00
0.25
0
200
400
600
800
1000
cor_item__Intercept[1]__Intercept[3]
400
0
0.0
0.5
1.0
r_item__Intercept[1]__Intercept[
1.0
0.0
0
200
400
600
800
1000
cor_item__Intercept[2]__Intercept[3]
400
0
0.00
0.25
0.50
0.75
1.00
r_item__Intercept[2]__Intercept[
1.00
0.00
0
200
400
600
800
1000


4000
3000
2000
1000
0
0.950
0.975
1.000
1.025
1.050
disc
1.050
1.025
1.000
0.975
0.950
disc
0
200
400
600
800
1000
Chain
1
2
3
4

PCM difficulty - ability map

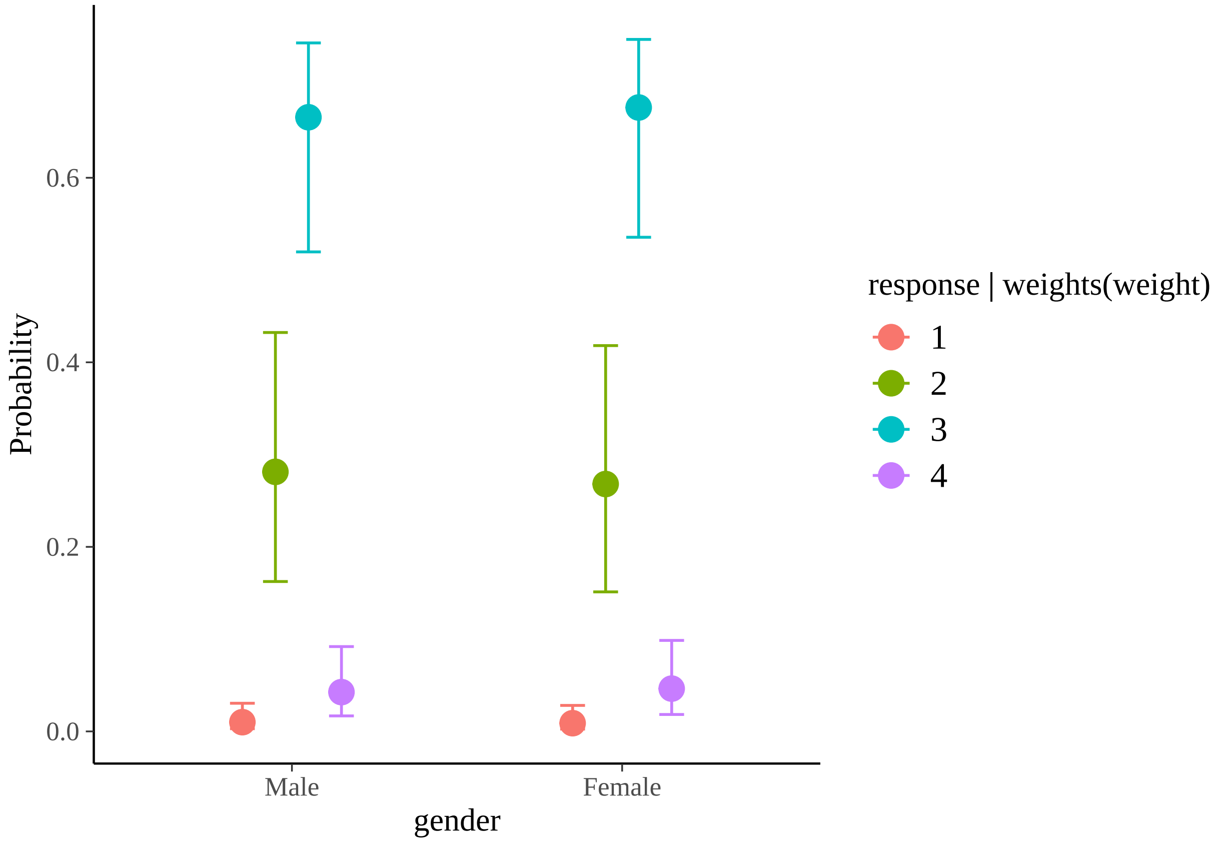


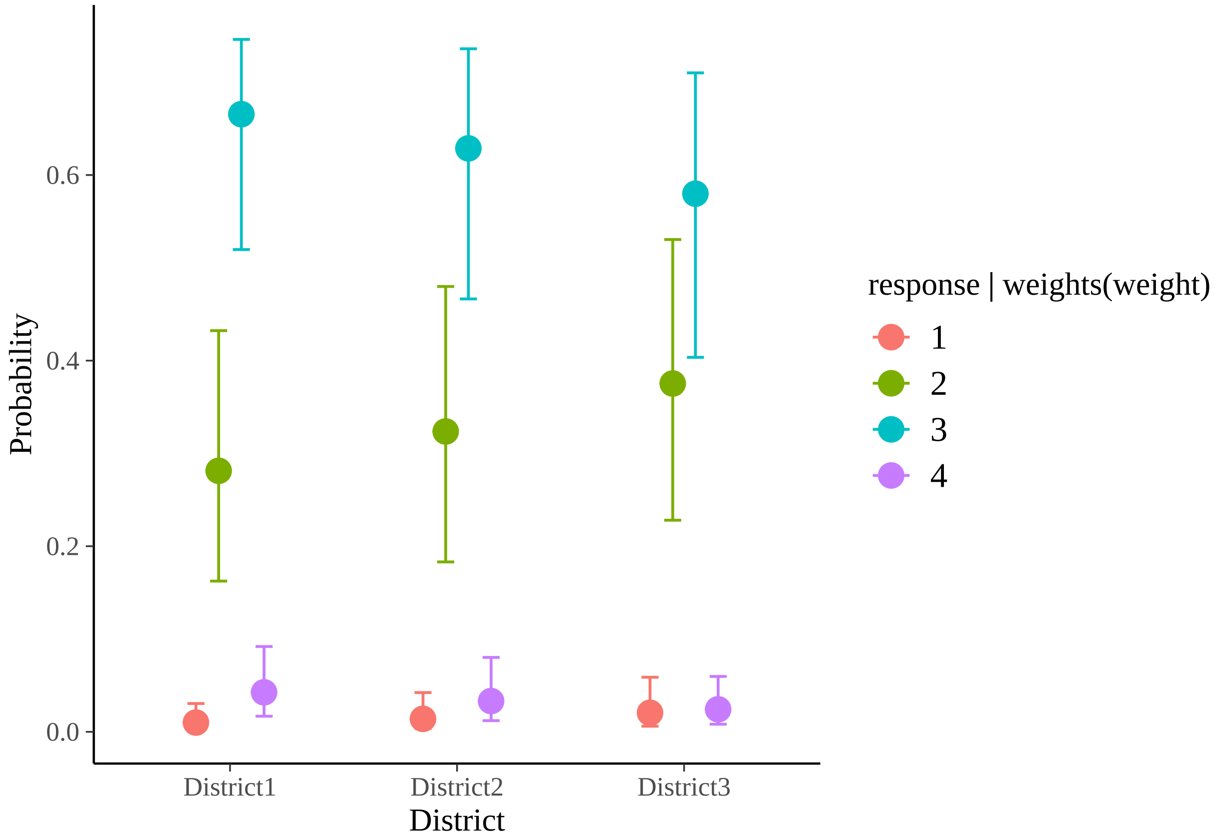
DIF

I used an approach that is almost the same as in the belief scale case – I included gender, year, and district as fixed effects, and then added gender and district to the random effects at the item level.

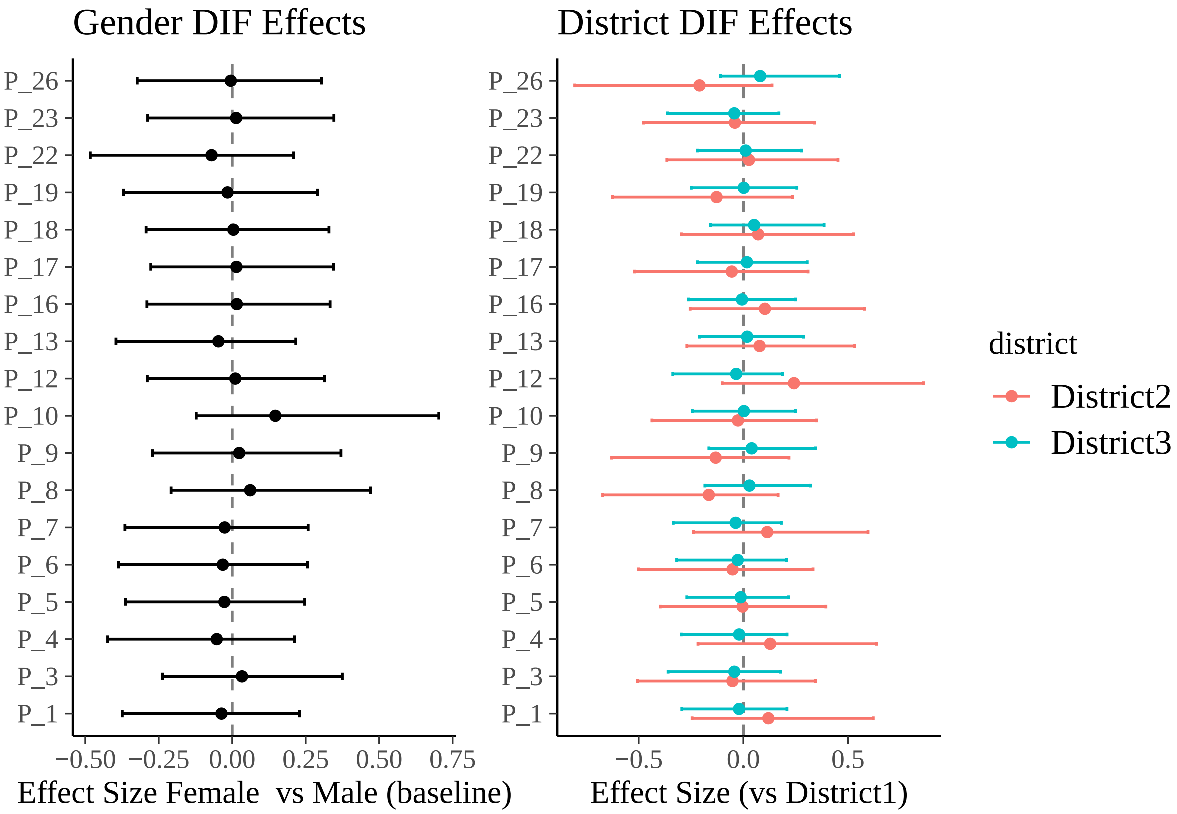
Uniform DIF

```
##  Family: acat 
##   Links: mu = logit; disc = identity 
## Formula: response | weights(weight) ~ gender + YearG + District + (1 || item) + (0 + gender + Distric
##    Data: long_data_dif (Number of observations: 2661) 
##   Draws: 4 chains, each with iter = 2000; warmup = 1000; thin = 1;
##          total post-warmup draws = 4000
## 
## Multilevel Hyperparameters:
## ~item (Number of levels: 18) 
##                         Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS
## sd(Intercept)               1.12      0.22     0.78     1.65 1.01     1482
## sd(genderMale)              0.14      0.10     0.01     0.38 1.00     1445
## sd(genderFemale)            0.25      0.13     0.02     0.52 1.00     1247
## sd(DistrictDistrict2)       0.22      0.14     0.01     0.51 1.00     1295
## sd(DistrictDistrict3)     0.11      0.08     0.00     0.31 1.00     2355
##                       Tail_ESS
## sd(Intercept)             2116
## sd(genderMale)            2311
## sd(genderFemale)          1394
## sd(DistrictDistrict2)     1679
## sd(DistrictDistrict3)     2200
##
## ~person (Number of levels: 151)
##               Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
## sd(Intercept)     1.07      0.08     0.92     1.23 1.00     1323     2142
##
## Regression Coefficients:
##                   Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
## Intercept[1]         -3.33      0.35    -4.03    -2.64 1.00      865     1440
## Intercept[2]         -0.86      0.34    -1.53    -0.19 1.00      815     1401
## Intercept[3]          2.75      0.34     2.09     3.44 1.00      865     1477
## genderFemale          0.07      0.20    -0.33     0.47 1.00     1352     1915
## YearGYear8           -0.05      0.22    -0.48     0.38 1.01     1021     1968
## YearGYear9           -0.05      0.25    -0.53     0.43 1.00     1203     1629
## DistrictDistrict2    -0.19      0.25    -0.69     0.30 1.00     1267     1927
## DistrictDistrict3    -0.42      0.23    -0.87     0.04 1.00     1161     1757
##
## Further Distributional Parameters:
##      Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
## disc     1.00      0.00     1.00     1.00   NA       NA       NA
##
## Draws were sampled using sampling(NUTS). For each parameter, Bulk_ESS
## and Tail_ESS are effective sample size measures, and Rhat is the potential
## scale reduction factor on split chains (at convergence, Rhat = 1).
```


Probability
0.6
0.4
0.2
0.0
Male
Female
gender
response | weights(weight)
1
2
3
4



The results show that there is almost no clear uniform DIF at the gender and district levels, and that people of different genders and districts have similar responses to the questionnaire. Except for a very marginal DIF tendency in district 3 -0.42 [–0.87, 0.04] , which suggests that teachers working in low SES schools may use connectionist pedagogic practice less frequently than teachers working in high SES schools. From the uniform DIF graph we can also see the pattern that respondents from district 3 are more likely to endorse category 2 and less likely to endorse category 3. While the evidence is not statistically strong enough to say it is a uniform DIF effect.



Item sepecific DIF

There is no substantial evidence of Differential Item Functioning (DIF) within the items as well, as the posterior distributions of group differences remain close to zero with narrow credible intervals. This suggests that the questionnaire responses were largely invariant across different respondent groups.