belief-practice_Chi_0122

Chi Zhang

2025-01-22

```
## Registered S3 method overwritten by 'GGally':
##   method from
##   +.gg   ggplot2
```

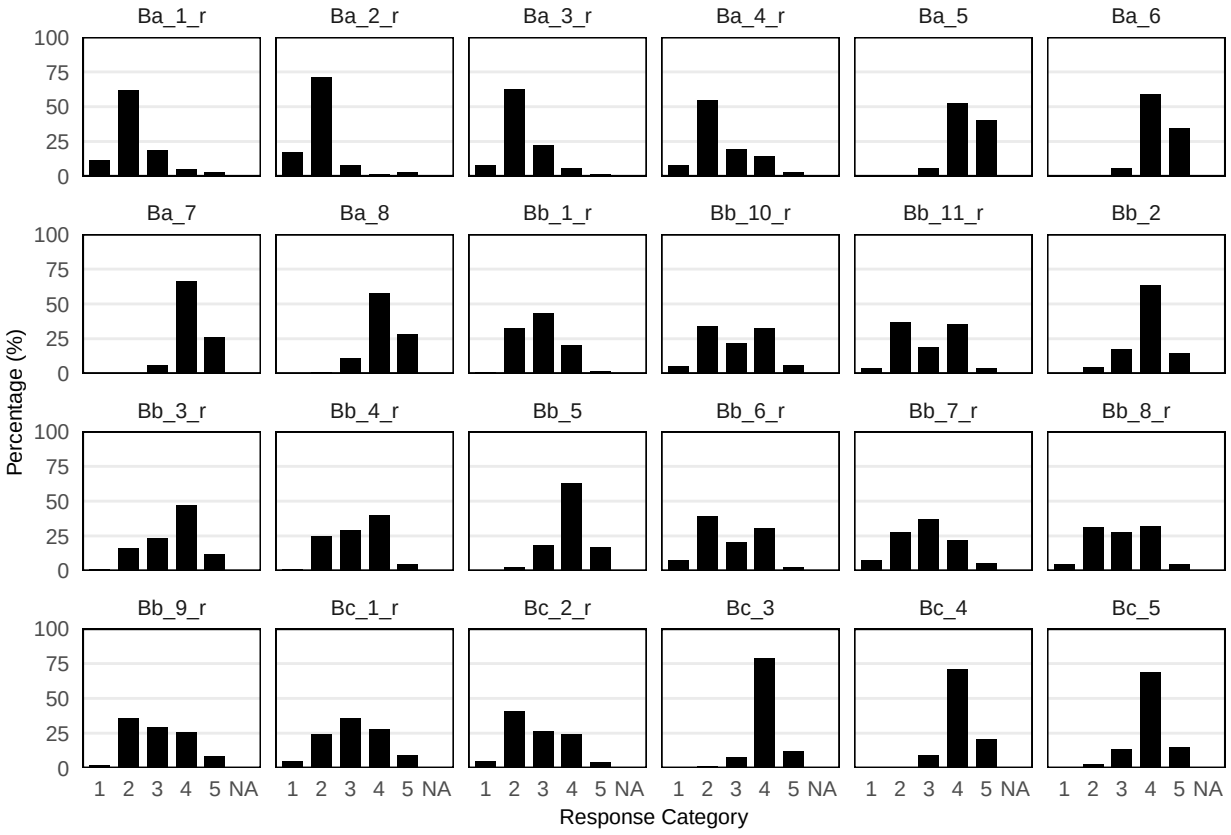
1	Mathematics is a collection of rules and procedures.
2	Mathematics involves remembering and applying definitions, formulas, mathematical facts and procedures.
3	To do mathematics requires much practice of correct application of routines.
4	When solving mathematical tasks you need to know the correct procedure in advance.
5	Mathematics involves creativity and new ideas.
6	Many aspects of mathematics have practical relevance.
7	Mathematics helps solve everyday problems and tasks.
8	In mathematics many things can be discovered and tried out by oneself.

1	Students must be able to solve mathematics problems quickly.
2	Students must be able to understand why the answer is correct.
3	Students learn mathematics best by attending to the teacher's standard explanations.
4	For students, non-standard procedures can interfere with learning the correct procedure.
5	Students can figure out a way to solve mathematical problems without a teacher's help.
6	Only the more able students can participate in mathematics activities.
7	To be good at mathematics one needs to be talented.
8	Among other forms of talent, mathematical talent is relatively fixed
9	Boys tend to be more talented in mathematics than girls.
10	Attention to mathematically weak students can interfere with the learning experience of other students.
11	Attention to mathematically talented students can interfere with the learning experience of other students.

1	Innovative pedagogy (e.g., project-based learning, flipped classroom) is not worth the time and expense.
2	Innovative pedagogy (e.g., project-based learning, flipped classroom) should only play a minor, supplementary role in the mathematics classroom.
3	I'm continuously seeking better ways to teach mathematics.
4	Teachers should help students using mathematics to critically analyze the world.
5	Discussion of social issues should be involved in the maths classroom.

Data Check

This dataset is a part of my survey project, which investigated belief and practice of secondary mathematics teacher in China. In this essay, I'll specifically look into belief instrument. It is about teachers belief about equity. The very basic assumption is the better performance of teachers, the more inclusive they are, while the others are more exclusive. While the insturment is designed a bit more complicated than that - it includes 24 items with 8 for belief about mathematics (start with Ba), 11 for belief about learners (start with Bb) and 5 for belief about pedagogy (start with Bc). There are some items are of reverse wording, I have already cleaned this by replace the original score with reverse score, which with a "_r" end. Let's first have a peek about the dataset, the response distribution, and missing data.



(Missing) Data cleaning

Gender&YearG (stands for yeargroup) will be used later in DIF. Here, we first just look at the Belief instrument, the 24 items. There is only 155 observations for this data, and the extreme categogies (1, and sometimes 5) are seldom appear. When looking into missing data in terms of items, there is some compeletely missing in category 1&2. There are possible two strategies to treat this: i. to collapse 1/2 for every item; ii. to collapse 1/2 for items which that have weak response in this two categories. While look through all the response distribution, I find 1/2 are often too thin to provide enough information, and if I choose ii, then it is problematic if I want to conduct RSM or GRM (and possibily any IRT without flexible threshold), becaosue they won't have any data about few thresholds; also considering the small sample nature of the dataset, a simple collapsing solution is easier to conduct, and possibly raising the reliability wihile not losing too much information. Also looking into missing data in terms of person - I want to keep as much information as possible instead of simple deleting them. Thus I look at those who missing more than 40% of the instruments (missing_count >= 10), and find out 3 persons miss the whole instrument. Of course we should delete them

- but how about others? Instead of treating them as trunky data and throw away, I want to try a weighted penalty solution. The goal is to avoid discarding participants with missing data outright while ensuring that those with significant gaps contribute less to the analysis. Here's how it works:

- **If someone is missing fewer than 10 items (roughly 40%)**, they receive a full weight of 1, meaning their responses are treated as completely reliable.
- **If someone is missing all 24 items**, their weight drops to 0, and they are excluded from the analysis.
- **For participants in between**, their weight decreases step by step according to an exponential decay function. The penalty grows more severe as the number of missing items increases, reflecting reduced confidence in their data.

Now based on the two strategies, let's try to fit them into simple models to see whether it works - basically, RSM in Beyesian is used as a foundation model, with 3 variations (i. null model, ii. collapsing model, iii. weighted collapsing model)

Let's see the comparasion of the different models:

```
## [1] "Comparison 1: Null RSM vs Weighted RSM"
```

```
## [1] "----------------------------------------"
```

```
## Output of model 'fit_null_rsm':
##
## Computed from 4000 by 3599 log-likelihood matrix.
##
##          Estimate    SE
## elpd_loo  -4042.1  51.1
## p_loo       142.8   4.4
## looic      8084.2 102.2
## ------
## MCSE of elpd_loo is 0.2.
## MCSE and ESS estimates assume MCMC draws (r_eff in [0.4, 2.2]).
##
## All Pareto k estimates are good (k < 0.7).
## See help('pareto-k-diagnostic') for details.
##
## Output of model 'fit_w_rsm':
##
## Computed from 4000 by 3599 log-likelihood matrix.
##
##          Estimate   SE
## elpd_loo  -4016.8 49.2
## p_loo       141.5  3.8
## looic      8033.5 98.5
## ------
## MCSE of elpd_loo is 0.2.
## MCSE and ESS estimates assume MCMC draws (r_eff in [0.5, 2.6]).
##
## All Pareto k estimates are good (k < 0.7).
## See help('pareto-k-diagnostic') for details.
##
## Model comparisons:
##              elpd_diff se_diff elpd_loo se_elpd_loo p_loo   se_p_loo looic
```

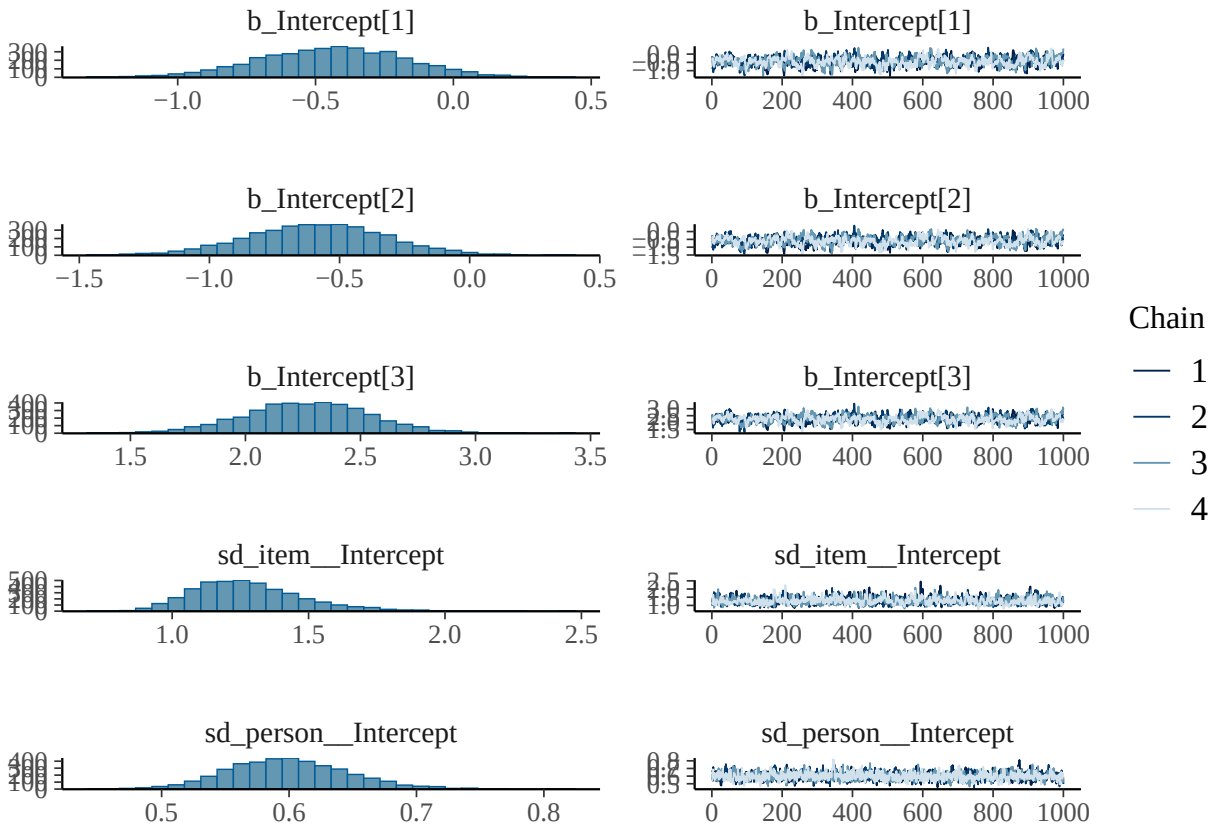
```
## fit_w_rsm        0.0       0.0 -4016.8     49.2       141.5     3.8   8033.5
## fit_null_rsm   -25.3       8.9 -4042.1     51.1       142.8     4.4   8084.2
##              se_looic
## fit_w_rsm       98.5
## fit_null_rsm   102.2

## [1] "\nComparison 2: Weighted RSM vs Weighted Collapsed RSM"

## [1] "----------------------------------------"

## Output of model 'fit_w_rsm':
##
## Computed from 4000 by 3599 log-likelihood matrix.
##
##          Estimate   SE
## elpd_loo  -4016.8 49.2
## p_loo       141.5  3.8
## looic      8033.5 98.5
## ------
## MCSE of elpd_loo is 0.2.
## MCSE and ESS estimates assume MCMC draws (r_eff in [0.5, 2.6]).
##
## All Pareto k estimates are good (k < 0.7).
## See help('pareto-k-diagnostic') for details.
##
## Output of model 'fit_w_c_rsm':
##
## Computed from 4000 by 3599 log-likelihood matrix.
##
##          Estimate   SE
## elpd_loo  -3551.2 44.8
## p_loo       144.9  3.7
## looic      7102.4 89.6
## ------
## MCSE of elpd_loo is 0.2.
## MCSE and ESS estimates assume MCMC draws (r_eff in [0.4, 2.2]).
##
## All Pareto k estimates are good (k < 0.7).
## See help('pareto-k-diagnostic') for details.
##
## Model comparisons:
##             elpd_diff se_diff elpd_loo se_elpd_loo p_loo   se_p_loo looic
## fit_w_c_rsm     0.0       0.0 -3551.2     44.8       144.9     3.7   7102.4
## fit_w_rsm    -465.6      25.3 -4016.8     49.2       141.5     3.8   8033.5
##             se_looic
## fit_w_c_rsm    89.6
## fit_w_rsm      98.5
```

It works! In the result, we can see the null model works the worst - when I introduce the weighted mechanism in, the model works better, and when I further collapse the 1/2 category it works significantly better. We can have a look at the best model (for now), fit_w_c_rsm (which is RSM with weighted/category collapsed response data).

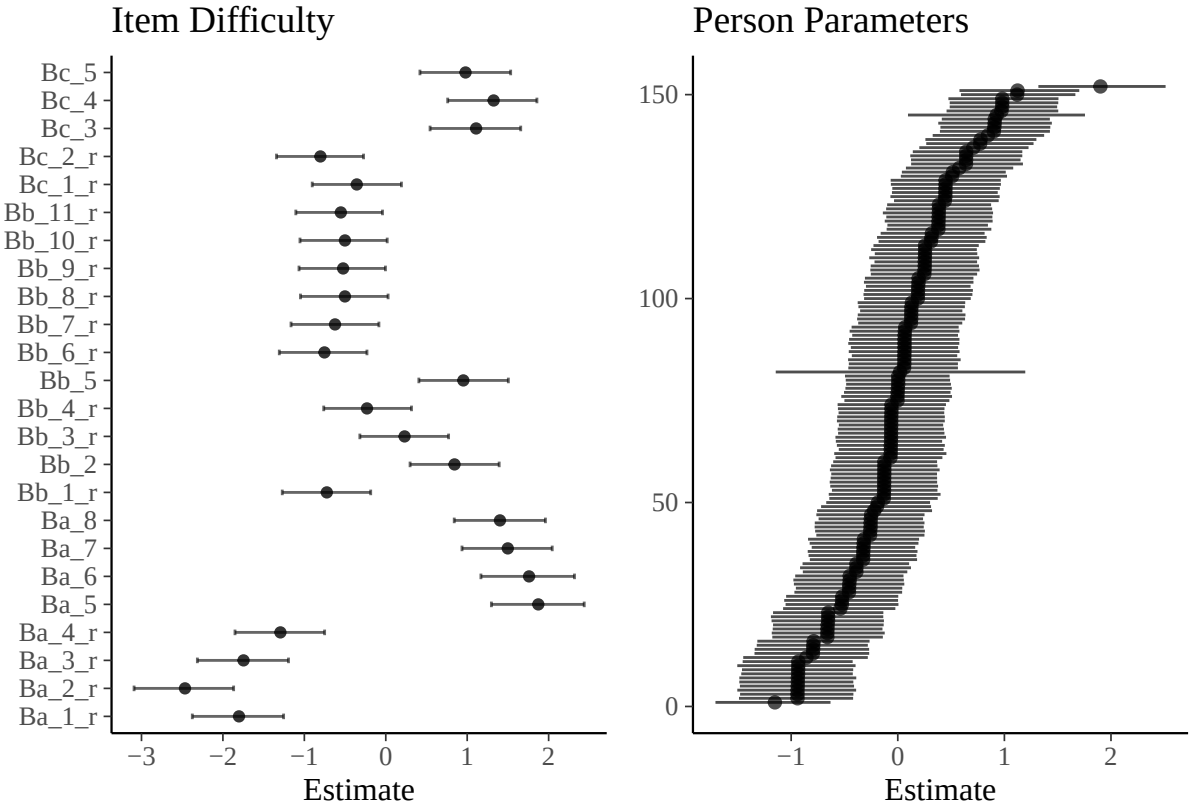
Simplist foundational Model check

Lets start by checking the fit statistics:

```
##  Family: acat
##   Links: mu = logit; disc = identity
## Formula: response | weights(weight) ~ 1 + (1 | item) + (1 | person)
##    Data: long_data_collapsed (Number of observations: 3599)
##   Draws: 4 chains, each with iter = 2000; warmup = 1000; thin = 1;
##          total post-warmup draws = 4000
##
## Multilevel Hyperparameters:
## ~item (Number of levels: 24)
##               Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
## sd(Intercept)     1.28      0.20     0.96     1.74 1.00      879     1719
##
## ~person (Number of levels: 152)
##               Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
## sd(Intercept)     0.60      0.05     0.51     0.70 1.00     1430     1906
##
## Regression Coefficients:
##              Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
## Intercept[1]    -0.44      0.27    -0.97     0.07 1.01      463     1093
## Intercept[2]    -0.59      0.27    -1.11    -0.07 1.01      445     1079
## Intercept[3]     2.27      0.27     1.72     2.79 1.01      474      994
##
## 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).
```



Also have a look at person/item map. For item difficulty, the model seems like have a quite good distribution; and though in person map there are few outliers or people with large HDI (indicating the uncertainty of the ability estimate), it is overall acceptable implying a good fit, which means we could maintain all the items (and participants) for now.



Now let's look into the RSM structure. It simply assumes that 24 items are unidimensional - they all contribute to one specific construct, which is mathematics teacher's belief about equity. While we can simply tell this is not true based on intuition (and theory) - people with general inclusive belief about education might hold something different when it takes the context of mathematics. Simply put, there might be different dimension of belief about equity, and these different dimension groups might (not) share the commoness (which can be perceived as general belief about equity).

Dimensionality

The original instrument already provided a guess based on theory - we may believe there are at least three dimensions, namely belief about mathematics, belief about learners, and belief about pedagogy. Now we want to know what is the latent construct of belief - how these three sub-beliefs connect with each other (or not)?

Potential multidimensional solution to iterate model

Here is three possible solutions with different model settings: 1. **Unidimensional** - we assume that belief about equity is unidimensional, which means three dimensions are co-linear (which we just proved is problematic);

2. **Correlated Traits model** - we assume the 24 item instrument contribute into 3 constructs(beliefs), and they may link with each other, while we don't assume there is a explicit general construct can explain all three;

3. **Bifactor multidimensional model** - we assume there is a general factor (general belief about equity) that affect all 24 items, while within each subinstruments there are specific factors/constructs (belief

about maths/learner/pedagogy) is situation specific, and is orthogonal and independent with general belief. G&S factors together explain the instrument performance.

i. The Bayesian package (brms) used in this study, to the best of my knowledge, does not yet support a simple and intuitive coding solution for directly fitting standard Bifactor Model. Given that brms is built on the GLMM framework, I opted for a GLMM-based approximation to the Bifactor structure that aligns better with the package's capabilities.

ii. I have also experimented with using more specialized tools like mirt to fit complex models (without Beyesian approach), including standard Bifactor models. However, the small sample size in this study resulted in sparse information, leading to difficulties in model convergence and increased parameter instability.

Competing MIRT choices BF/CT

Let's first try the two choices. Note that for unidimensional rsm model, the model is rather simple and of less parameter pressure, thus it is easy to use the default settings for converge. While for complex models with more parameters to estimate, we often adapt settings to make model better converge. This is at expense of calculation pressure, often cost more time and computational power. Two common strategies to do this is i. adding iteration times, e.g., from 2000 to 4000; ii. adding control settings, e.g., adapt_delta = 0.99, max_treedepth = 15. Solely choosing ii is a more economical option for me here.

model comparasion (null vs ct vs bf)

```
##             elpd_diff se_diff elpd_loo se_elpd_loo p_loo   se_p_loo looic
## fit_ct_rsm      0.0       0.0 -3486.7     46.1       270.7     7.2   6973.5
## fit_bf_rsm     -1.1       1.9 -3487.9     46.2       267.5     7.2   6975.7
## fit_w_c_rsm   -64.5      13.1 -3551.2     44.8       144.9     3.7   7102.4
##             se_looic
## fit_ct_rsm     92.2
## fit_bf_rsm     92.3
## fit_w_c_rsm    89.6
```

model summary (bf ct)

```
##  Family: acat
##   Links: mu = logit; disc = identity
## Formula: response | weights(weight) ~ 1 + (1 | item) + (1 | person) + (0 + dima || person) + (0 + di
##    Data: long_data_collapsed (Number of observations: 3599)
##   Draws: 4 chains, each with iter = 2000; warmup = 1000; thin = 1;
##          total post-warmup draws = 4000
##
## Multilevel Hyperparameters:
## ~item (Number of levels: 24)
##               Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
## sd(Intercept)     1.40      0.23     1.06     1.94 1.00      837     1398
##
## ~person (Number of levels: 152)
##               Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
## sd(Intercept)     0.57      0.06     0.46     0.69 1.00     1597     2415
## sd(dima)          0.53      0.10     0.35     0.72 1.00     1124     1548
```

```
## sd(dimb)          0.58      0.07     0.44     0.72 1.00     1610     2371
## sd(dimc)          0.33      0.13     0.05     0.57 1.01      466      672
##
## Regression Coefficients:
##              Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
## Intercept[1]    -0.53      0.29    -1.12     0.05 1.01      418      917
## Intercept[2]    -0.53      0.29    -1.10     0.04 1.01      399      857
## Intercept[3]     2.48      0.29     1.88     3.06 1.01      420     1050
##
## 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).

##  Family: acat
##   Links: mu = logit; disc = identity
## Formula: response | weights(weight) ~ 1 + (1 | item) + (0 + dima + dimb + dimc | person)
##    Data: long_data_collapsed (Number of observations: 3599)
##   Draws: 4 chains, each with iter = 2000; warmup = 1000; thin = 1;
##          total post-warmup draws = 4000
##
## Multilevel Hyperparameters:
## ~item (Number of levels: 24)
##               Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
## sd(Intercept)     1.38      0.22     1.03     1.91 1.00      855     1498
##
## ~person (Number of levels: 152)
##                Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
## sd(dima)           0.73      0.08     0.58     0.89 1.00     2298     2809
## sd(dimb)           0.82      0.07     0.70     0.96 1.00     1888     2711
## sd(dimc)           0.72      0.09     0.56     0.91 1.00     2109     2916
## cor(dima,dimb)     0.37      0.11     0.15     0.57 1.00      778     1749
## cor(dima,dimc)     0.52      0.12     0.26     0.75 1.00     1297     2117
## cor(dimb,dimc)     0.65      0.09     0.46     0.81 1.00     2192     2827
##
## Regression Coefficients:
##              Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
## Intercept[1]    -0.56      0.28    -1.13    -0.04 1.02      347      940
## Intercept[2]    -0.56      0.28    -1.10    -0.04 1.02      335      905
## Intercept[3]     2.46      0.28     1.89     2.99 1.02      345      854
##
## 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).
```

Statistically, the multidimensional choice is significantly better than the null model (elpd_diff is way larger than 2*se_diff), and the CT-RSM works more or less the same with BF-RSM. While there is one problem in

both two models (and actually also in the null rsm model)- the intercept [1] & [2] are almost identical. Recall my category collapse move, we shall know that now category 1 is (strongly) disagree, 2 is neutral, 3 is agree, 4 is strongly agree. Thus intercept [1] & [2] stands for shifts from disagree to neutral, and neutral to agree. Thus the same of intercept [1]&[2] means the model sees little distinction between respondents choosing "neutral" versus "agree"/"(strongly) disagree". Statistically, one simple approach is to further collapse the categories - to add 'neutral' into 'agree' or into "(strongly) disagree"; while linguistically and practically, this is unacceptable. For the original category collapsing move, the two categories disagree and strongly disagree more or less points to same 'negative' direction, but now adding neutral to either agree or disagree category will cause explanation difficulties; also, comparing to the first collapse move where the data don't have enough observation in category 1&2 to estimate, we do have plenty of observations people endorsing neutral. We shall find a way to make sense of it.

Model iteration (1PL vs 2PL)

we already discussed that the model needs a multidimensional structure (though not sure bf/ct which is better). Now, I want to try to make intercept [1] & [2] less identical - at least the "neutral" can be as a statistically proper category. One approach is to introduce the discrimination parameter, considering how good each items can distinguish participants (item with higher discrimination means that people with higher ability are more likely to score higher, that being 'distinguished' among others). We take structure BF/CT together, using GRSM (the 2pl version of RSM) for iteration test.

Threshold check

```
##  Family: acat
##   Links: mu = logit; disc = log
## Formula: response | weights(weight) ~ 1 + (1 | i | item) + (1 | person) + (0 + dima || person) + (0 +
##          disc ~ 1 + (1 | i | item)
##    Data: long_data_collapsed (Number of observations: 3599)
##   Draws: 4 chains, each with iter = 2000; warmup = 1000; thin = 1;
##          total post-warmup draws = 4000
##
## Multilevel Hyperparameters:
## ~item (Number of levels: 24)
##                               Estimate Est.Error l-95% CI u-95% CI Rhat
## sd(Intercept)                     5.00      1.43     2.59     8.08 1.00
## sd(disc_Intercept)                0.76      0.12     0.56     1.03 1.00
## cor(Intercept,disc_Intercept)     0.89      0.05     0.76     0.96 1.00
##                               Bulk_ESS Tail_ESS
## sd(Intercept)                      792     1184
## sd(disc_Intercept)                 825     1695
## cor(Intercept,disc_Intercept)      913     2046
##
## ~person (Number of levels: 152)
##               Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
## sd(Intercept)     1.14      0.34     0.57     1.88 1.00      689     1240
## sd(dima)          1.77      0.51     0.89     2.86 1.00      787     1279
## sd(dimb)          1.20      0.36     0.59     2.01 1.00      723     1314
## sd(dimc)          0.91      0.34     0.34     1.63 1.00      445      320
##
## Regression Coefficients:
##                Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
## Intercept[1]      -0.40      0.97    -2.49     1.31 1.01      331      564
```

```
## Intercept[2]       0.08      0.93    -1.89     1.76 1.01      322      469
## Intercept[3]       5.10      1.45     2.49     8.27 1.01      530      764
## disc_Intercept    -0.62      0.29    -1.14     0.00 1.00      625     1145
##
## 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 + (1 | i | item) + (0 + dima + dimb + dimc | person)
##          disc ~ 1 + (1 | i | item)
##    Data: long_data_collapsed (Number of observations: 3599)
##   Draws: 4 chains, each with iter = 2000; warmup = 1000; thin = 1;
##          total post-warmup draws = 4000
##
## Multilevel Hyperparameters:
## ~item (Number of levels: 24)
##                               Estimate Est.Error l-95% CI u-95% CI Rhat
## sd(Intercept)                     4.68      1.40     2.26     7.83 1.00
## sd(disc_Intercept)                0.79      0.12     0.57     1.06 1.00
## cor(Intercept,disc_Intercept)     0.89      0.05     0.76     0.96 1.00
##                               Bulk_ESS Tail_ESS
## sd(Intercept)                      758     1384
## sd(disc_Intercept)                 874     1820
## cor(Intercept,disc_Intercept)     1071     2213
##
## ~person (Number of levels: 152)
##                 Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
## sd(dima)            1.93      0.57     0.95     3.17 1.00      736     1267
## sd(dimb)            1.45      0.43     0.71     2.36 1.00      699     1227
## sd(dimc)            1.38      0.41     0.67     2.29 1.00      770     1303
## cor(dima,dimb)      0.30      0.10     0.11     0.48 1.01      523     1383
## cor(dima,dimc)      0.49      0.10     0.28     0.67 1.00      944     1663
## cor(dimb,dimc)      0.47      0.10     0.26     0.66 1.00     1427     2288
##
## Regression Coefficients:
##                 Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
## Intercept[1]       -0.20      0.89    -2.12     1.41 1.02      267      635
## Intercept[2]        0.22      0.86    -1.59     1.83 1.02      278      562
## Intercept[3]        4.77      1.42     2.31     7.81 1.01      699     1663
## disc_Intercept     -0.54      0.30    -1.08     0.13 1.01      856     1735
##
## 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).
```

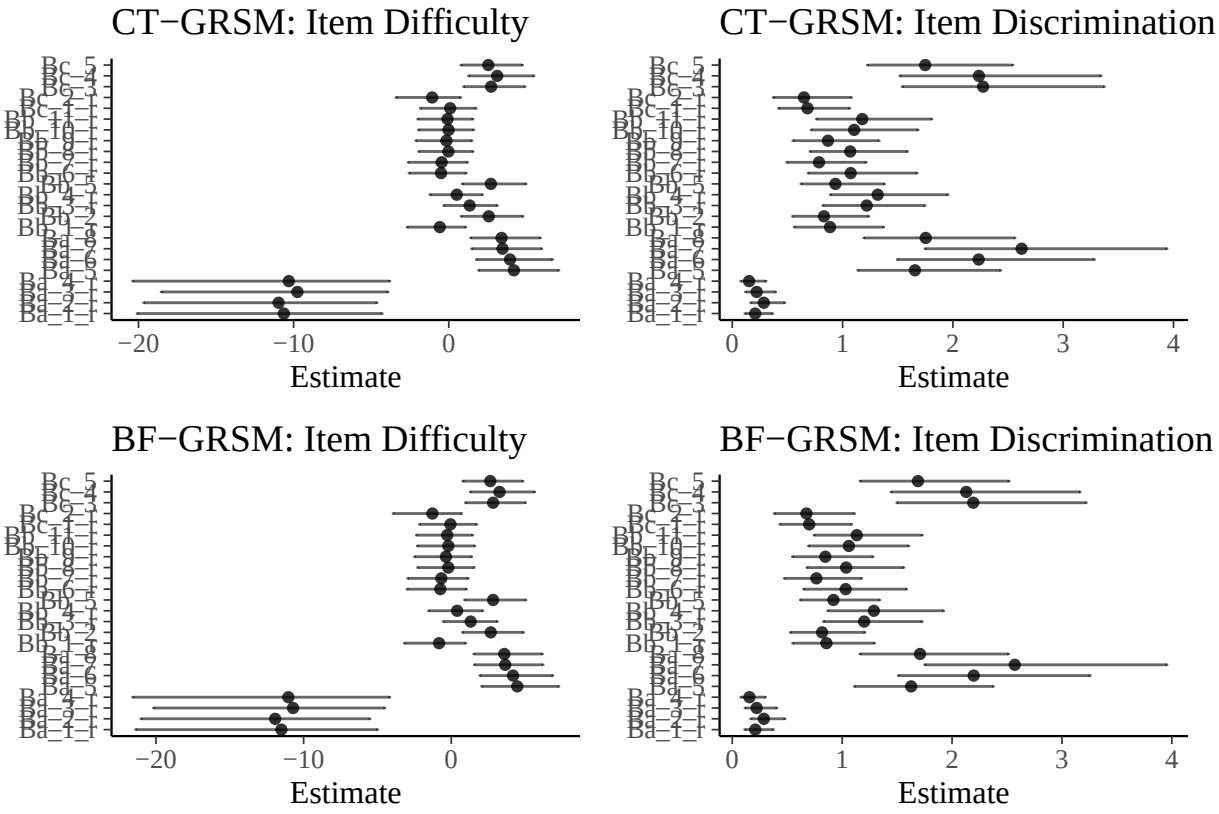
The summary shows when we introduce the discrimination parameter in, the models obtains more information to distinguish intercept 1/2. And the reason is quite obvious - in 1pl, we assume every item with same discrimination (=1). This essentially forces the slope of the latent trait curve to be identical across all items, which means that no matter how a respondent's latent ability (or attitude) varies, the model cannot adjust how sharply each item shifts between response categories.

By contrast, once we allow each item to have its own discrimination parameter (as in a 2PL approach), an item with substantial empirical evidence to differentiate adjacent categories—such as "Strongly Disagree"

vs. “Neutral,” or “Neutral” vs. “Agree”—can exhibit a steeper or shallower slope. And we can clearly see that cor(Intercept,disc_Intercept) = 0.89 [0.75, 0.96], which means disc is highly associate with difficulty (which make the assumption of every disc = 1 problematic).

item fit check

Now lets check the item difficulty/disc map.



Some really interesting results here. Firstly, as we already find out, ct and bf functions similarly well - we see the performance of items in two models is quite similar as well. Secondly, we can see ba_1_r ~ ba_4_r (reversed items) works statistically really ‘bad’. it shows shows extremely ‘hard’ that everyone scores low. In addition, the reason why these items were not found to be abnormal in the previous item check as ‘outliers’ is that the assumed discrimination=1 absorbed a lot of abnormal information that should have been reported. Here, the discrimination is extremely low, means that the items can’t distinguish people from ability well - not only respondents can easily score high, but also respondents perform nearly randomly, regardless their ability.

Let’s look at the items:

1	Mathematics is a collection of rules and procedures.
2	Mathematics involves remembering and applying definitions, formulas, mathematical facts and procedures.
3	To do mathematics requires much practice of correct application of routines.
4	When solving mathematical tasks you need to know the correct procedure in advance.

These are all classic items being used and validated in similar research/instrument, while the results shows they work unwell for my sample. Some possible explanations might be:

i) These items were designed to assess teachers' beliefs about the nature of mathematics on a relatively linear spectrum, with one end representing a rigorous, structured view full of rules and procedures, and the other end representing a creative, exploratory perspective. However, in these items, this contrast is not explicitly articulated, which might have led our sample to fail to perceive the intended meaning of the items.

ii) This could also be due to certain cultural differences. For example, teachers in Western countries might interpret these items with a "yes, but" approach, opting for more conservative responses as a way of reconciling the tension between creativity and structure. In contrast, Chinese teachers may view these two aspects as independent—believing that mathematics can simultaneously be both creative and procedural. This cultural difference could explain why these items did not perform well in our sample.

iii) Lastly, this might indicate the multidimensional nature even within the belief about mathematics. While these items were intended to assess a single dimension, their extreme ease likely made them incapable of detecting the complexity of this construct, resulting in their inability to differentiate respondents across ability levels.

With the reasons being said, the four items are deleted due to poor functionality.

Model Iteration (Competing MIRT with different link function)

With the new item set and 2pl approach, I'll re-construct competing models. Models with different link function - RSM/GRM/PCM - are tested. Based on its multidimensional structure, and 2pl choice, 6 models are constructed as bf-grsm, bf-grm, bf-gpcm; ct-grsm, ct-grm, ct-gpcm.

Loo check

```
## [1] "Model comparison (LOO):"
```

```
##                 elpd_diff se_diff elpd_loo se_elpd_loo p_loo   se_p_loo looic
## fit_ct_gpcm_new     0.0       0.0 -2811.9     41.2       363.7    11.6   5623.9
## fit_bf_gpcm_new    -8.6       5.0 -2820.5     41.2       361.0    11.2   5641.0
## fit_bf_grm_new    -32.1      10.6 -2844.1     40.2       334.4    10.1   5688.2
## fit_ct_grm_new    -35.2      10.5 -2847.1     40.2       337.6    10.2   5694.2
## fit_ct_grsm_new   -40.6       9.4 -2852.5     40.6       322.7    10.5   5705.1
## fit_bf_grsm_new   -42.9       9.5 -2854.9     40.6       324.3    10.6   5709.7
##                 se_looic
## fit_ct_gpcm_new    82.3
## fit_bf_gpcm_new    82.4
## fit_bf_grm_new     80.5
## fit_ct_grm_new     80.5
## fit_ct_grsm_new    81.1
## fit_bf_grsm_new    81.2
```

```
## [1] "\nPareto k diagnostic summary:"
```

```
##     Model High_K Total_Obs Low_K
## 1 CT_GRSM      9      2993  2984
```

```
## 2  CT_GRM      2      2993  2991
## 3 CT_GPCM     13      2993  2980
## 4 BF_GRSM     10      2993  2983
## 5  BF_GRM      1      2993  2992
## 6 BF_GPCM     17      2993  2976
```

```
## [1] "\nP_loo to parameters ratio:"
```

```
##     Model    P_loo Parameters Ratio
## 1 CT_GRSM 322.6680        203 1.589
## 2  CT_GRM 337.6274        203 1.663
## 3 CT_GPCM 363.6785        203 1.792
## 4 BF_GRSM 324.2650        203 1.597
## 5  BF_GRM 334.4326        203 1.647
## 6 BF_GPCM 361.0047        203 1.778
```

According to the loo_check, we can see that CT/BF with same IRT structure still performance pretty similar, while for models of differnet linking function, GPCM works the best, followed by GRM, with GRSM at the end. This is also the order of P_loo, which suggest that more parameter is used for estimating the model, the more information it provides. While the downside is, more parameters will cause overfitting risk and the instability of the model. The best model GPCMs (in LOO comparasion), are detected with 17/13 high K observations in BF/CT structure, considering the P-loo > p, this is a potential sign for over-fitting. GRM performance clearly better in terms of stability, with 1 high K in BF-GRM, 2 in CT-GRM. Considering the major advantage - flexible threshold setting in PCM - is not necessary for my following analysis, I would prefer to give up a moderate amount of model information in PCM, for better stability in GRM, especially considering the difference in performance between the two models is not that large, with approximately 3*se_diff.

Multidimensional Structure Check (ECV/FC)

Due to we can't determine which multidimensional structure is better, I use further index to check their multidimensional assumption.

ECV Explained Common Variance is used to check BF - it simply means how good the general latent trait can used to explain the sub-dimensions. We expect a not too high yet not too low score, means that general belief do make sense, but for each dimension, they have their own important sub-belief to explain the others. FC is used to check CT (factor correlations) - it represents the correlations between different dimension-specific factors.

Results shows that both works - FC among three CT models shows that three dimensions related with each other at a moderate level; ECV among three BF models shows that there is a general belief that could explain moderate but significant information of different dimensions.

Considering the stability of GRM, and the theoretical compatibility (CT technically rejects the existence of general belief yet ECV proves the rationale of general belief in BF structure), I decided to take BF structure into final model.

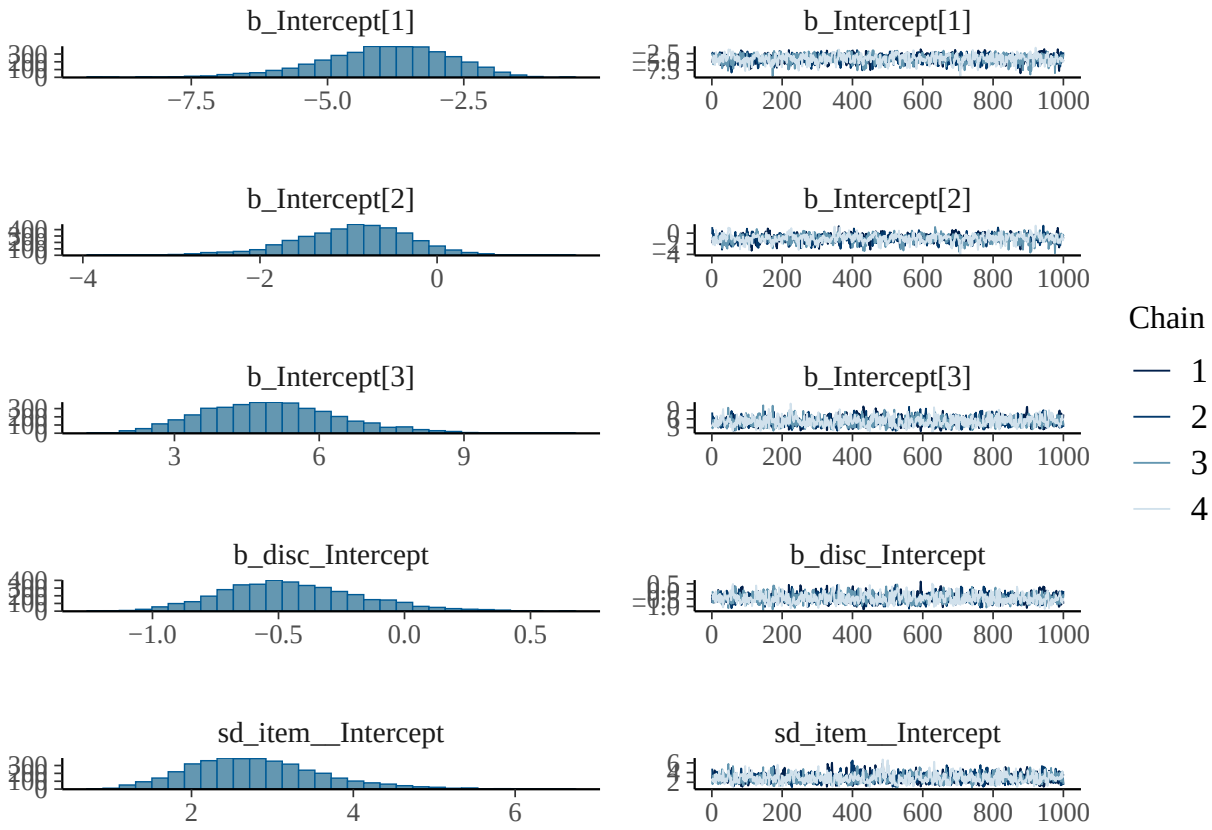
```
##     Model                ECV_A                ECV_B                ECV_C
## 1 CT-GRSM                 <NA>                 <NA>                 <NA>
## 2  CT-GRM                 <NA>                 <NA>                 <NA>
## 3 CT-GPCM                 <NA>                 <NA>                 <NA>
## 4 BF-GRSM 0.282 [0.163, 0.398] 0.486 [0.337, 0.662] 0.618 [0.378, 0.870]
## 5  BF-GRM 0.318 [0.190, 0.442] 0.450 [0.297, 0.605] 0.626 [0.378, 0.855]
## 6 BF-GPCM 0.153 [0.051, 0.282] 0.610 [0.313, 0.876] 0.571 [0.290, 0.897]
```

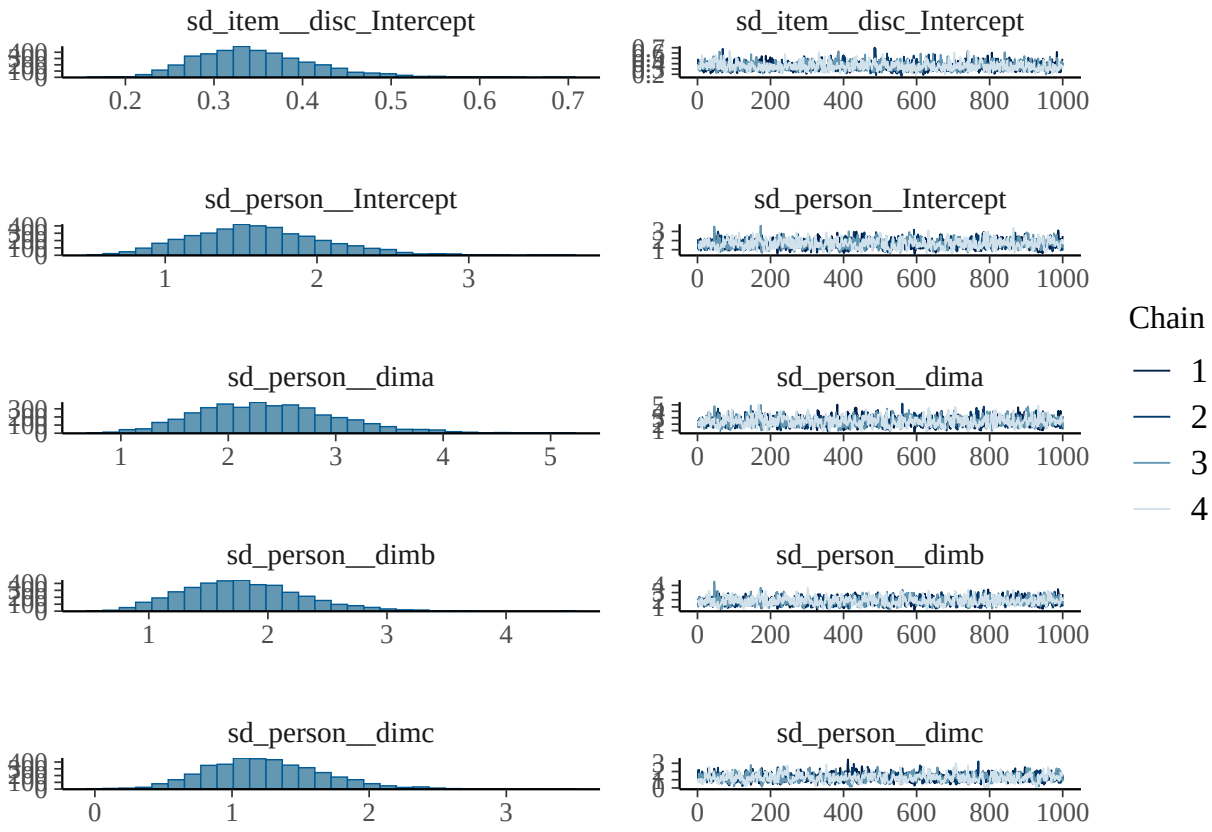
```
##                   FC_AB                FC_AC                FC_BC
## 1 0.293 [0.111, 0.481] 0.509 [0.311, 0.683] 0.469 [0.286, 0.673]
## 2 0.315 [0.131, 0.491] 0.491 [0.300, 0.663] 0.463 [0.262, 0.645]
## 3 0.243 [0.054, 0.421] 0.486 [0.262, 0.676] 0.445 [0.224, 0.667]
## 4                 <NA>                 <NA>                 <NA>
## 5                 <NA>                 <NA>                 <NA>
## 6                 <NA>                 <NA>                 <NA>
```

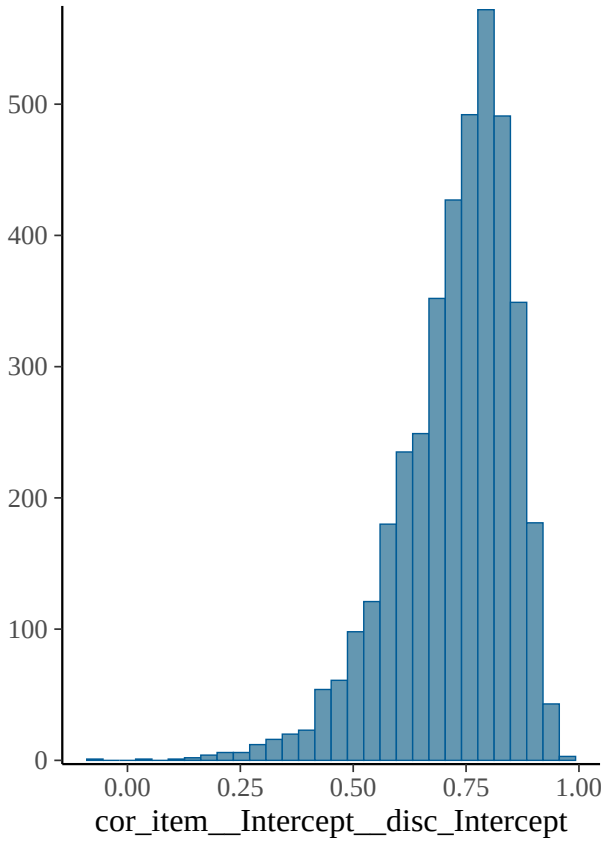
Final model choice

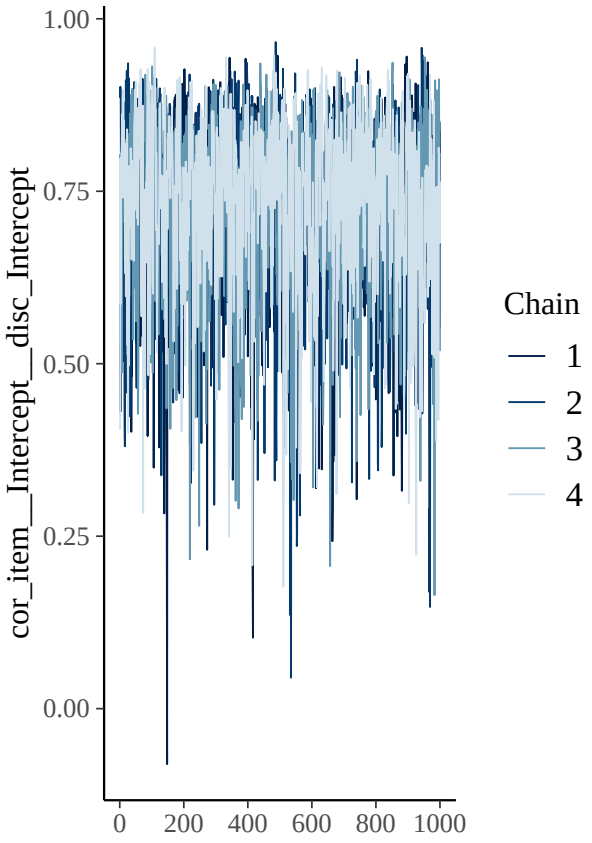
I finally choose to use bf structure, with 2pl approach, GRM linking function, thus combining bf-grm model.
Here is the mmodel summary and the fit plot.

```
##  Family: cumulative
##   Links: mu = logit; disc = log
## Formula: response | weights(weight) ~ 1 + (1 | i | item) + (1 | person) + (0 + dima || person) + (0 +
##          disc ~ 1 + (1 | i | item)
##    Data: long_data_collapsed_new (Number of observations: 2993)
##   Draws: 4 chains, each with iter = 2000; warmup = 1000; thin = 1;
##          total post-warmup draws = 4000
##
## Multilevel Hyperparameters:
## ~item (Number of levels: 20)
##                               Estimate Est.Error l-95% CI u-95% CI Rhat
## sd(Intercept)                     2.80      0.84     1.39     4.69 1.00
## sd(disc_Intercept)                0.35      0.07     0.24     0.50 1.00
## cor(Intercept,disc_Intercept)     0.73      0.13     0.42     0.91 1.00
##                               Bulk_ESS Tail_ESS
## sd(Intercept)                      956     1185
## sd(disc_Intercept)                1296     2742
## cor(Intercept,disc_Intercept)     1394     2074
##
## ~person (Number of levels: 151)
##               Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
## sd(Intercept)     1.61      0.44     0.84     2.55 1.00      940     1270
## sd(dima)          2.38      0.66     1.26     3.85 1.00     1119     1477
## sd(dimb)          1.79      0.50     0.94     2.88 1.00     1148     1526
## sd(dimc)          1.25      0.43     0.49     2.16 1.00      875      952
##
## Regression Coefficients:
##                Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
## Intercept[1]      -3.96      1.19    -6.60    -1.93 1.00      866     1230
## Intercept[2]      -0.96      0.67    -2.46     0.24 1.00      725     1368
## Intercept[3]       5.03      1.42     2.58     8.11 1.00     1046     1512
## disc_Intercept    -0.44      0.28    -0.93     0.16 1.00     1099     1435
##
## 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).
```

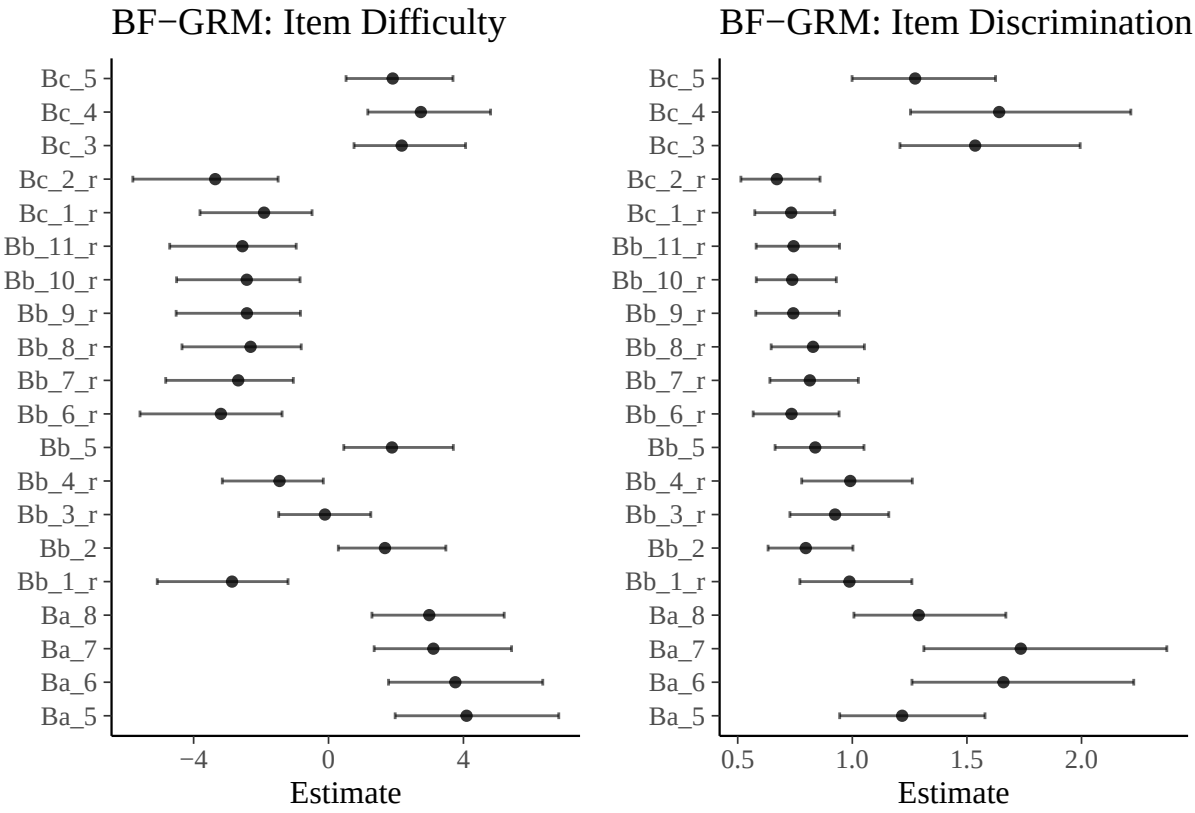

b_Intercept[1]
b_Intercept[1]
b_Intercept[2]
b_Intercept[2]
b_Intercept[3]
b_Intercept[3]
b_disc_Intercept
b_disc_Intercept
sd_item__Intercept
sd_item__Intercept
Chain
1
2
3
4


sd_item__disc_Intercept
sd_item__disc_Intercept
sd_person__Intercept
sd_person__Intercept
sd_person__dima
sd_person__dima
sd_person__dimb
sd_person__dimb
sd_person__dimc
sd_person__dimc
Chain
1
2
3
4

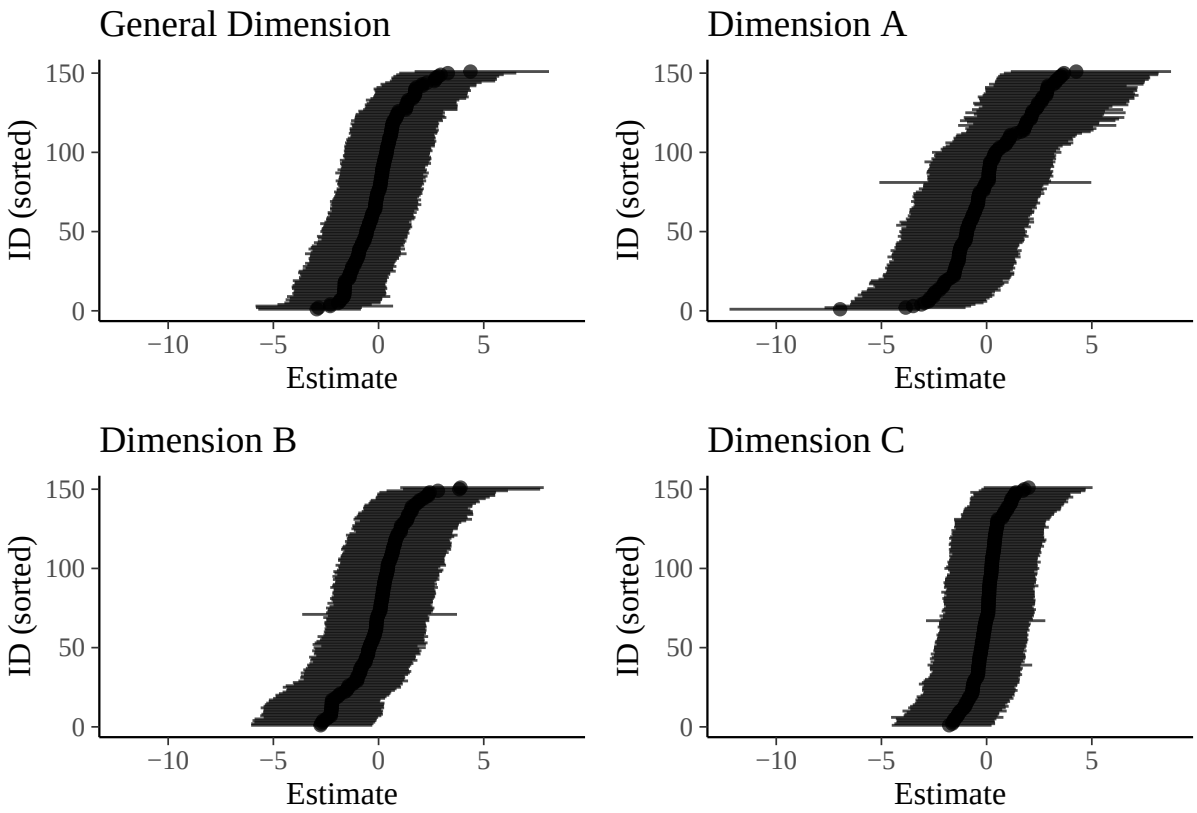

500
400
300
200
100
0
0.00
0.25
0.50
0.75
1.00
cor_item__Intercept__disc_Intercept


1.00
0.75
0.50
0.25
0.00
cor_item__Intercept__disc_Intercept
0
200
400
600
800
1000
Chain
1
2
3
4

difficulty-discrimation map

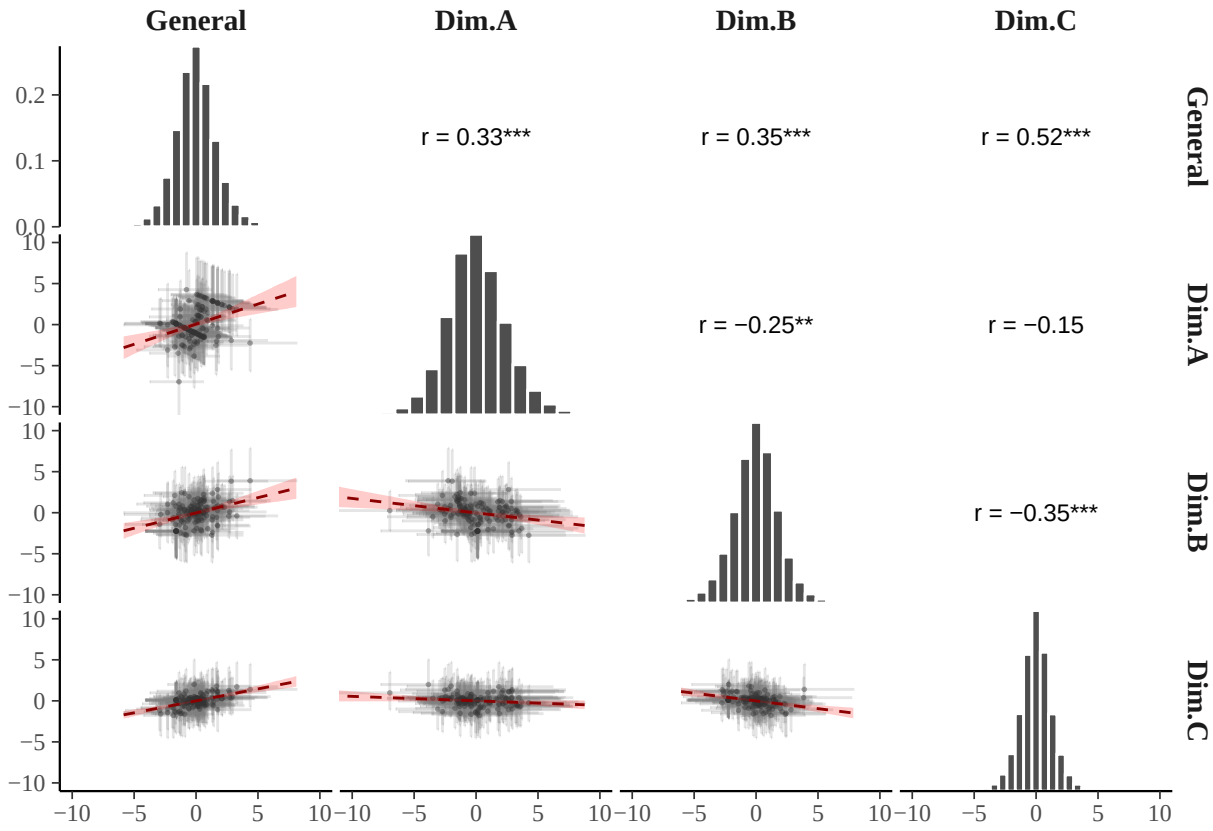


ability map with general & dimension specific beliefs



We can see several improvement in the iterations. Now we have clear three category thresholds, nice convergence (from good R hat and controlled k value), reasonable discrimination/difficulty/multidimensional ability distribution. For the abilities - DimC shows the greatest stability, with the narrowest posterior intervals and minimal variability across individuals, indicating consistent beliefs about pedagogy within the group. General Belief and DimB are moderately stable, with slightly wider posterior intervals and minor individual-level uncertainty in DimB. In contrast, DimA displays the highest uncertainty, characterized by larger standard deviations across individuals, broader posterior intervals, and a few outliers with significantly divergent posterior distributions. This variability may be due to the questionnaire structure—after removing simpler items (A1–A4), the remaining DimA items are concentrated around higher difficulty levels (approximately +4), leading to less precise estimates for respondents. Additionally, beliefs about mathematics may inherently vary more across individuals and be less susceptible to group-level patterns, resulting in greater randomness in estimates.

Posterior-Mean Associations Across Dimensions



In the bifactor structure, the General Factor and DimA, DimB, DimC are constrained to be orthogonal, meaning that, for example, DimA represents beliefs about mathematics after controlling for general inclusive beliefs. This plot shows simple regressions exploring associations between latent variables based on posterior mean estimates from 155 respondents. Overall, the results indicate the presence of a general inclusive belief: teachers who generally hold more inclusive beliefs tend to adopt more inclusive positions across all dimensions, with the strongest influence observed on beliefs about pedagogy (DimC), while the effects on beliefs about mathematics (DimA) and beliefs about learners (DimB) are moderate.

Interestingly, there is no statistical association between DimA and DimC, potentially reflecting the idea that teachers' beliefs about mathematics are less connected to their pedagogical beliefs. This could indicate that mathematical beliefs are perceived as more rigid or independent from pedagogical considerations.

Furthermore, DimB (beliefs about learners) shows moderate negative associations with both DimA and DimC, which might suggest a practical inclusive standpoint: teachers who strongly believe in equal opportunities for all learners to succeed in mathematics may, on the one hand, adopt a pragmatic view of schooling mathematics—seeing it as rigorous, binary (right or wrong), and procedural (DimA); on the other hand, they may place less emphasis on pedagogical innovation or view such innovation as secondary to ensuring equal access and opportunities (DimC). These associations suggest that a focus on inclusivity for learners might involve trade-offs in how teachers approach mathematical rigor and pedagogy.

Model DIF on Gender&District

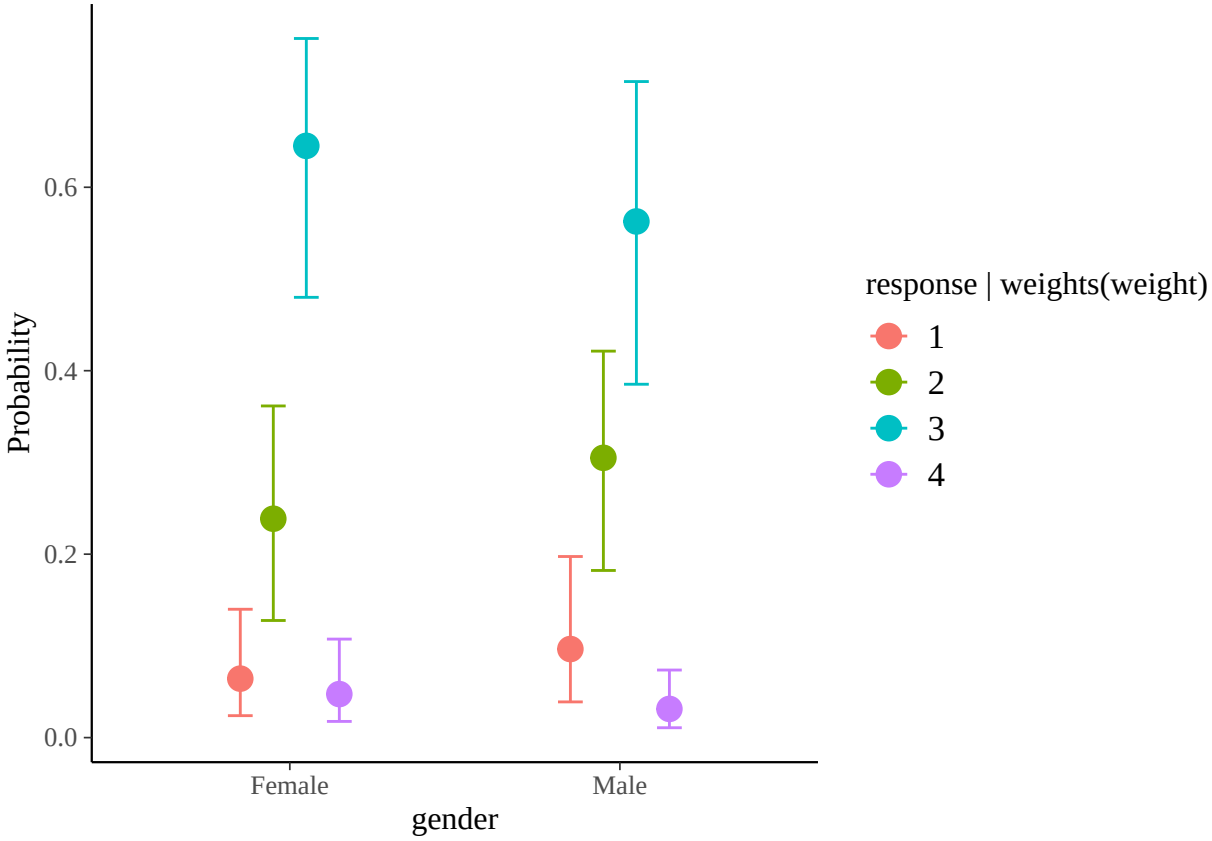
```
##  Family: cumulative
##   Links: mu = logit; disc = log
```

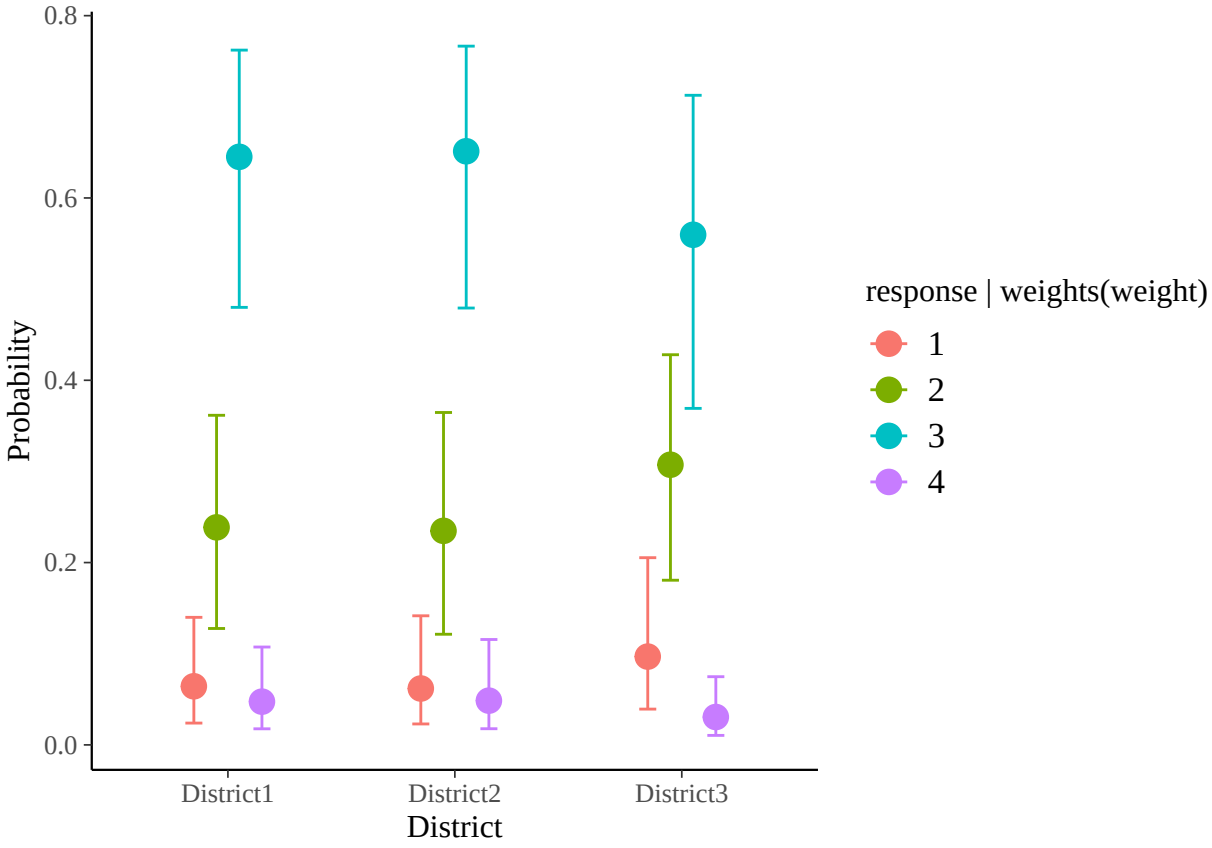
```
## Formula: response | weights(weight) ~ gender + YearG + District + (1 || item) + (0 + gender + Distri
##          disc ~ 1 + (1 | item)
##    Data: long_data_dif_new (Number of observations: 2913)
##   Draws: 4 chains, each with iter = 2000; warmup = 1000; thin = 1;
##          total post-warmup draws = 4000
##
## Multilevel Hyperparameters:
## ~item (Number of levels: 20)
##                       Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS
## sd(Intercept)             4.87      1.23     2.77     7.57 1.00      933
## sd(genderFemale)          0.50      0.37     0.02     1.44 1.00     1421
## sd(genderMale)            0.53      0.38     0.03     1.42 1.00     1492
## sd(DistrictDistrict2)     0.66      0.46     0.03     1.72 1.00     1441
## sd(DistrictDistrict3)     0.44      0.35     0.01     1.27 1.00     1727
## sd(disc_Intercept)        0.35      0.07     0.23     0.51 1.01     1149
##                       Tail_ESS
## sd(Intercept)             1357
## sd(genderFemale)          1942
## sd(genderMale)            2559
## sd(DistrictDistrict2)     2051
## sd(DistrictDistrict3)     1758
## sd(disc_Intercept)        1448
##
## ~person (Number of levels: 147)
##               Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
## sd(Intercept)     2.86      0.70     1.64     4.39 1.00      771     1435
## sd(dima)          4.49      1.06     2.64     6.75 1.00      792     1303
## sd(dimb)          3.41      0.84     2.00     5.20 1.00      771     1268
## sd(dimc)          2.46      0.72     1.22     4.05 1.00      816     1117
##
## Regression Coefficients:
##                   Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
## Intercept[1]         -8.00      2.07   -12.40    -4.40 1.00      773     1211
## Intercept[2]         -2.46      1.30    -5.24    -0.05 1.00      843     1466
## Intercept[3]          8.99      2.45     4.78    14.32 1.01      701     1108
## disc_Intercept       -1.07      0.24    -1.50    -0.58 1.01      717     1178
## genderMale           -1.30      0.71    -2.80    -0.04 1.00     2030     2436
## YearGYear8           -0.21      0.71    -1.69     1.16 1.00     2341     2627
## YearGYear9           -0.10      0.79    -1.68     1.43 1.00     2216     2359
## DistrictDistrict2     0.07      0.78    -1.54     1.61 1.00     2326     2396
## DistrictDistrict3    -1.37      0.76    -2.97    -0.04 1.00     1822     2095
##
## 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).
```

Based on the final model, I iterate a DIF version to include gender/yeargroup/district in fixed effects, and gender/districts in item random effects. The results shows their is a marginal uniform DIF on gender - that controlling other variables, male will generally performance 1.30 digit (log-odd scale, [-2.8, -0.04]) worse than female. While the upper bound of CI is nearly zero (-0.04), suggesting a highly uncertainty; Also a marginal uniform DIF on district - that the district of the least SES (district 3) perform 1.37 digit (log-odd scale, [-2.97, -0.04]) worse than district 1 (baseline, of highest SES). The DIF actually make sense - that male tends to hold a more exclusive belief (resonate with literature); while those teachers from lower SES areas are also with more exclusive beliefs, might because they are facing greater pressure to competition and lack

the necessary resources and support for inclusion. *Though these findings are marginal.*

Visualising uniform DIF





The differences on the log scale are relatively harder to interpret, while these figures provide a more intuitive visualization of how teachers from different genders and districts endorse various response categories. From the first figure, we can observe that male teachers are less likely than female teachers to endorse higher categories (3 and 4, representing "agree" and "strongly agree") and more likely to select lower categories (1 and 2, representing "disagree" and "neutral"). A similar tendency is observed in the second figure, where teachers from District 3, associated with lower SES, show a lower probability of endorsing higher categories and a higher probability of selecting lower categories compared to teachers from Districts 1 and 2, which are associated with higher SES.

Item-level DIF

While in item-level, the marginal uniform DIF is faded by various items. None of the items showed significant group differences, and all the posteriors of group differences for the items spanned 0. Still, there are some items worth further investigation for gender differences, though here we don't have enough statistical evidence to assert: 1. bc_5 & ba_7 are the two items with the strongest male advantage tendency, and they also point to the discussion of practical problems in mathematics classes – a kind of contextualisation of mathematics. This may indicate the advantage and tendency of male teachers in this regard. 2. bb_9 and bb_3 are the two items with the strongest female advantage tendencies. Specifically, bb_9 addresses gender aptitude stereotypes, where female teachers are more likely to disagree with or challenge the notion that mathematical ability is inherently tied to gender. This suggests that female teachers might be more aware of or sensitive to the negative impacts of such stereotypes and more inclined to foster an equitable learning environment. Similarly, bb_3, which relates to the use of non-standard methods for solving mathematical problems, highlights female teachers' openness to diverse problem-solving approaches.

Though these are quite interesting tendencies, the model does not provide strong enough evidence to draw definitive conclusions.