World Health Organization

Rasch: Measurement Assumptions

Dr. Carolina Fellinghauer

External Consultant WHO

Rasch Analysis

A serie of assumptions have to be tested. If the scale ratings comply to these assumptions, the total score is interval-scaled.

- Stochastic ordering (fit of data to model)
- Monotonicity (ordering or response options)
- No local response dependencies or LID (no significant correlations between items)
- Unidimensionality (one latent construct)
- No differential item functioning or DIF (no sample subgroup effects)

Rasch Analysis

Procedure:

1) Estimation of the Item Difficulty Parameter
2) Estimation of the Person Ability Parameter
3) Obtaining the Residual Matrix:

   Residual Matrix: standardised difference for the observed ratings and the expected ratings based on the estimated the person ability and item difficulty parameter.

   Allows to test the Measurement Properties.

   The Residual Matrix should be free of any patterns.

Rasch Analysis

A serie of assumptions have to be tested. If the scale ratings comply to these assumptions, the total score is interval-scaled.

- Stochastic ordering (fit of data to model)
- Monotonicity (ordering or response options)
- No local response dependencies or LID (no significant correlations between items)
- Unidimensionality (one latent construct)
- No differential item functioning or DIF (no sample subgroup effects)

Infit and Outfit

Fit of items to the data is given in R-package eRm with the infit and outfit statistics.

Other statistics exist – Chi-square, F-test.

```
> itemfit(person.parameter(PCM.model))

Itemfit Statistics:
    Chisq df p-value Outfit MSQ Infit MSQ Outfit t Infit t Discrim
I1 18.689 19   0.477      0.934     0.987   -0.141   0.033   0.160
I2 24.268 19   0.186      1.213     1.110    0.771   0.493  -0.008
I3 19.352 19   0.434      0.968     1.041    0.006   0.248   0.104
I4 13.553 19   0.809      0.678     0.754   -1.147  -1.040   0.539
I5 14.376 19   0.761      0.719     0.786   -0.686  -1.009   0.504
I6 14.303 19   0.766      0.715     0.813   -0.723  -0.744   0.598
I7 21.986 19   0.285      1.099     1.048    0.447   0.267   0.234
```

Infit and Outfit

Fit of items to the data is given in R-package eRm with the infit and outfit statistics.

Other statistics exist – Chi-square, F-test.

```
> itemfit(person.parameter(PCM.model))

Itemfit Statistics:
    Chisq df p-value Outfit MSQ Infit MSQ Outfit t Infit t Discrim
I1 18.689 19   0.477      0.934     0.987   -0.141   0.033   0.160
I2 24.268 19   0.186      1.213     1.110    0.771   0.493  -0.008
I3 19.352 19   0.434      0.968     1.041    0.006   0.248   0.104
I4 13.553 19   0.809      0.678     0.754   -1.147  -1.040   0.539
I5 14.376 19   0.761      0.719     0.786   -0.686  -1.009   0.504
I6 14.303 19   0.766      0.715     0.813   -0.723  -0.744   0.598
I7 21.986 19   0.285      1.099     1.048    0.447   0.267   0.234
```

Infit and Outfit

To find the item fit requires computation of:

1) Expected response for each observation X_{ij}

$$E_{ij} = \Sigma_{k=0}^{mi} k(P_{ikj})$$

2a) The score residual Y_{ij}

$$Y_{ij} = X_{ij} - E_{ij}$$

2b) The standardized residual Z_{ij}

$$Z_{ij} = \frac{Y_{ij}}{(W_{ij})^{1/2}}$$

The variance of X_{ij} is formalized as

$$W_{ij} = \Sigma_{k=0}^{mi} (k - E_{ij})^2 P_{ikj}$$

Infit and Outfit

3) A chi-square statistic by summing the standardized residuals.

$$\chi^2 = \Sigma_{n=1}^{N} Z_{ij}^2$$

The chi-square divided by the sample size corresponds to the Mean-Square Outfit Statistic.

$$Outfit_i = \frac{\Sigma_{n=1}^{N} Z_{ij}^2}{N}$$

The Outfit Statistic is sensitive to outlier. To diminish the effect of outlier, the standardized residuals can be adjusted by their variance. This is the Mean- Square Infit Statistic.

$$Infit_i = \frac{\Sigma_{n=1}^{N} W_{ij} Z_{ij}^2}{\Sigma_{n=1}^{N} W_{ij}}$$

Underfit and Overfit

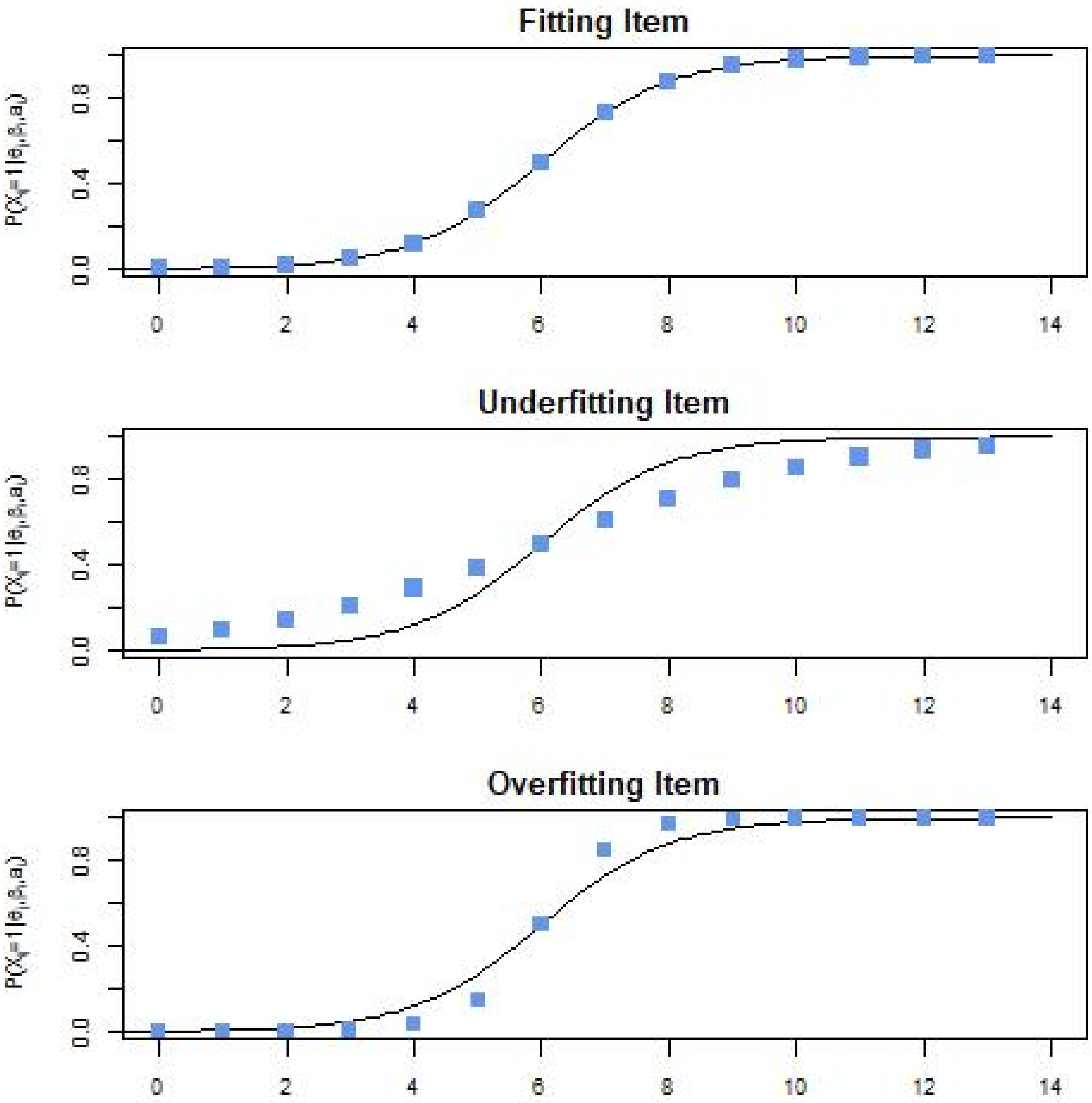
An item is fitting the Rasch model if the Infit and Outfit statistics are close to 1.

Underfit indicates underdiscrimination, the information is «blurred». It is not possible to differentiate ability levels. Underfit is found when the Infit or Outfit are much above 1.

Overfit indicates overderdiscrimination, the information is too «sharp». An overdiscriminating item acts like an on-off switch. Overfit is found when the Infit or Outfit are much below 1.

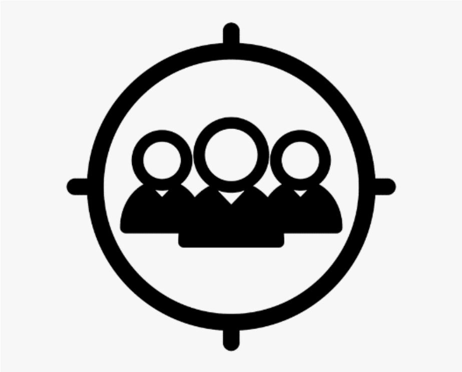
Note:

- Overfit is less critical for scales than underfit.
- Cut-off for acceptable fit, in terms of how much underfit can be tolerated, depends on the purpose of a scale.


Fitting Item
0.0
0.4
0.8
0
2
4
6
8
10
12
14
Underfitting Item
0.0
0.4
0.8
0
2
4
6
8
10
12
14
Overfitting Item
0.0
0.4
0.8
0
2
4
6
8
10
12
14

Targeting

Targeting indicates the degree to which the study population is outside the target range of the scale items



Targeting

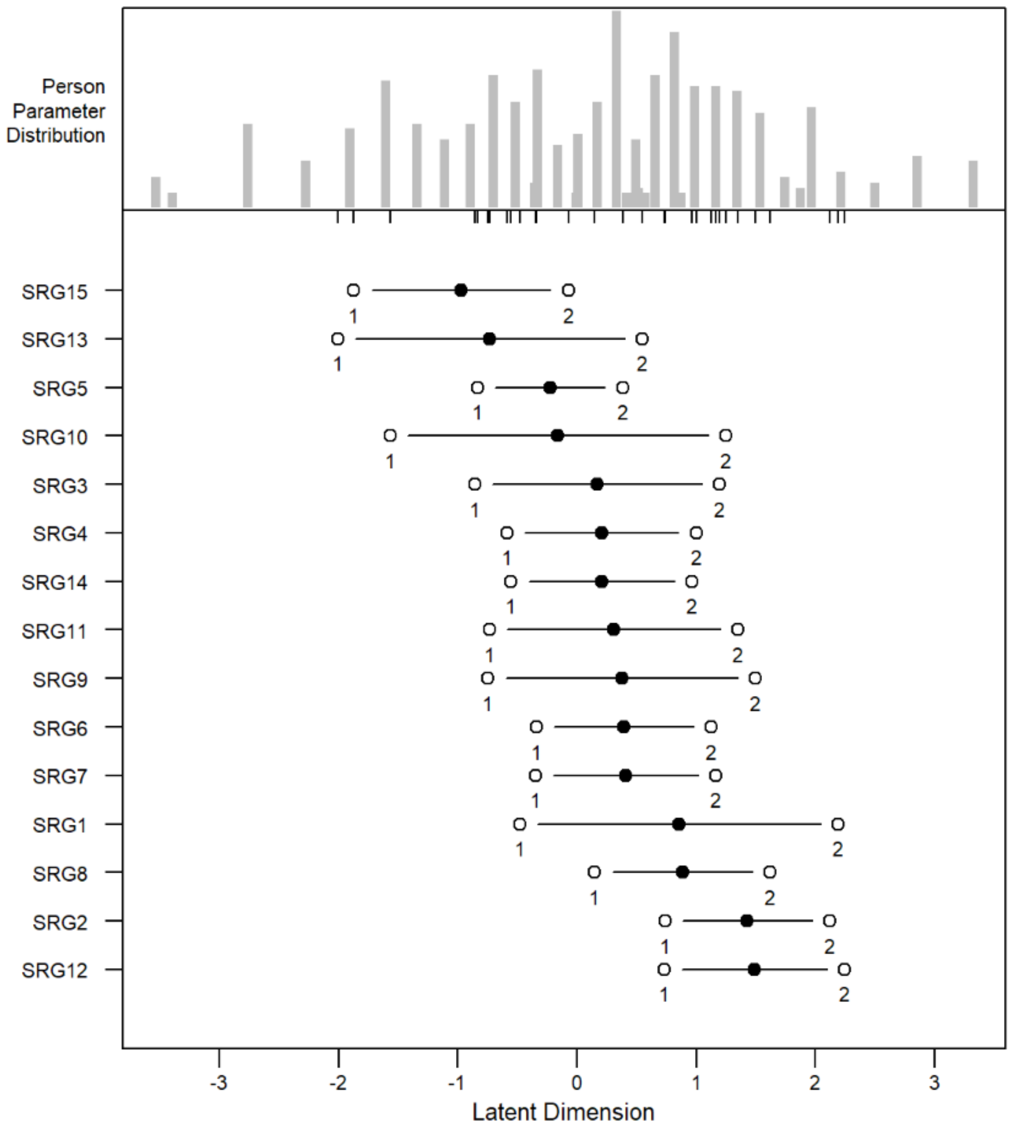
Item Difficulties approximate the person abilities

Characteristic of a well-targeted scale:

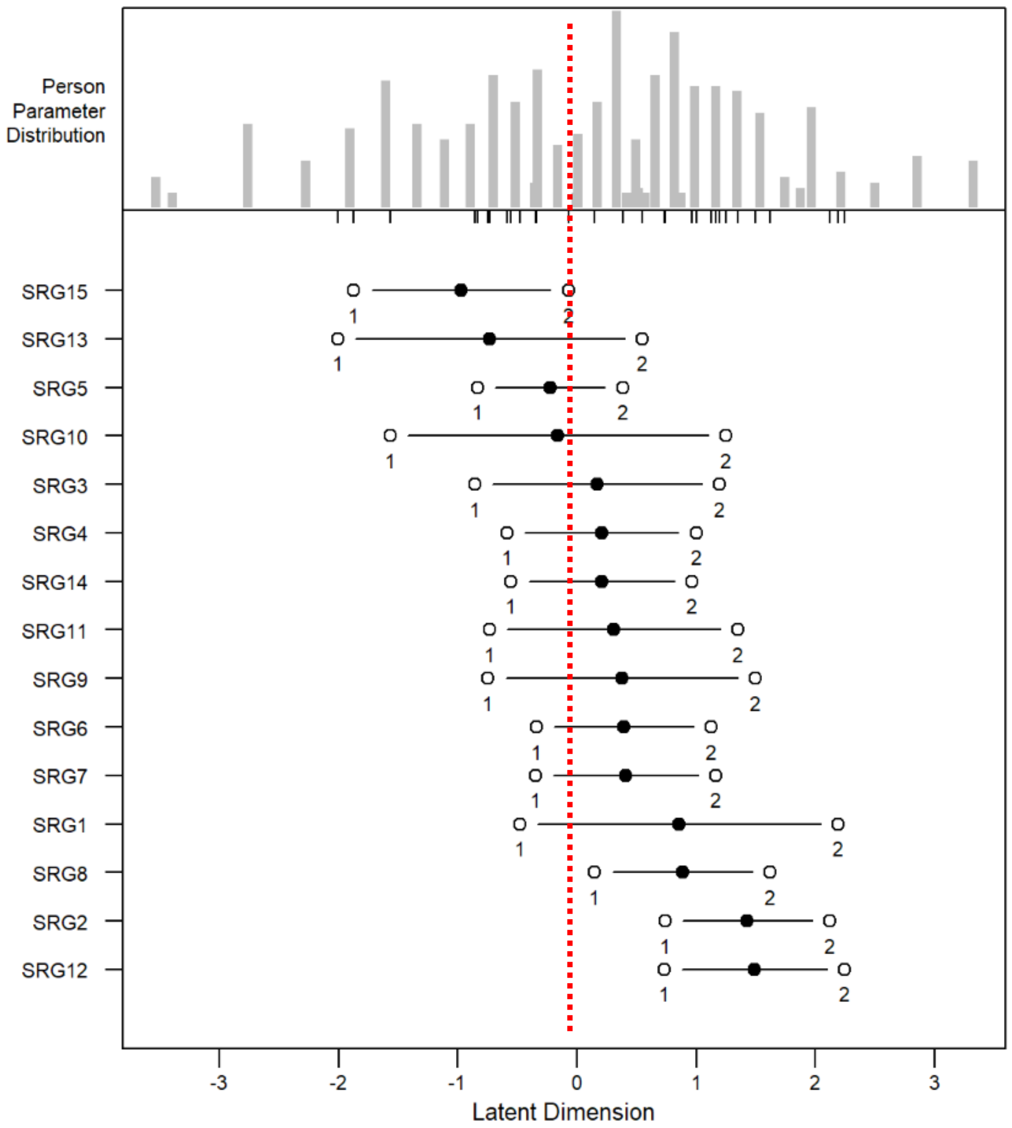
Difference mean difficulty and mean ability < 1 logit.

The SD of the item difficulty < 2.5
The SD of the person ability < 2.5

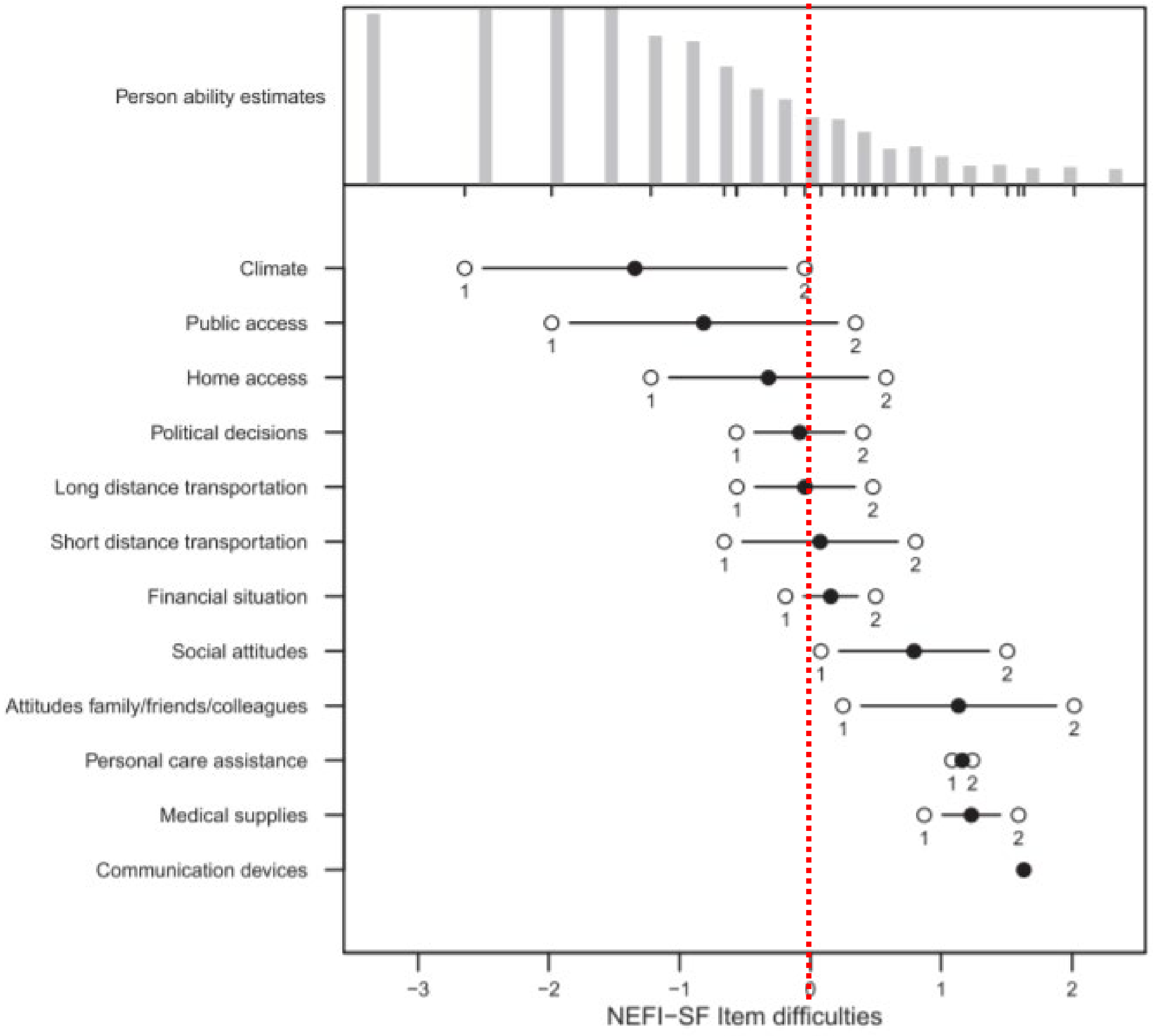
Person Item Map



Person Item Map



Person Item Map



Reliability

In the context of Modern Test Theory, reliability is a function of the variability and precision of the person ability estimates.

The Person Separation Reliability (PSR), calculates the proportion of person variance that is not due to error.

$$PSR = 1 - \left[\frac{MSE_p}{SD_p^2}\right]$$

MSE : Mean Square Person Measure Error
SD: The sample person measure variance

```
> SepRel(person.parameter(PCM.model))

Separation Reliability: 0.517
```

Reliability

The PSR ranges between 0 and 1.

PSR > 0.9 :
very good reliability, scale can be used for individual measurement

PSR > 0.85
good reliability, scale can be used for measurement at population level.

PSR > 0.7
low, but just sufficient reliability

PSR < 0.7
Insufficient reliability, scale cannot differentiate levels of abilities.

Rasch Analysis

A serie of assumptions have to be tested. If the scale ratings comply to these assumptions, the total score is interval-scaled.

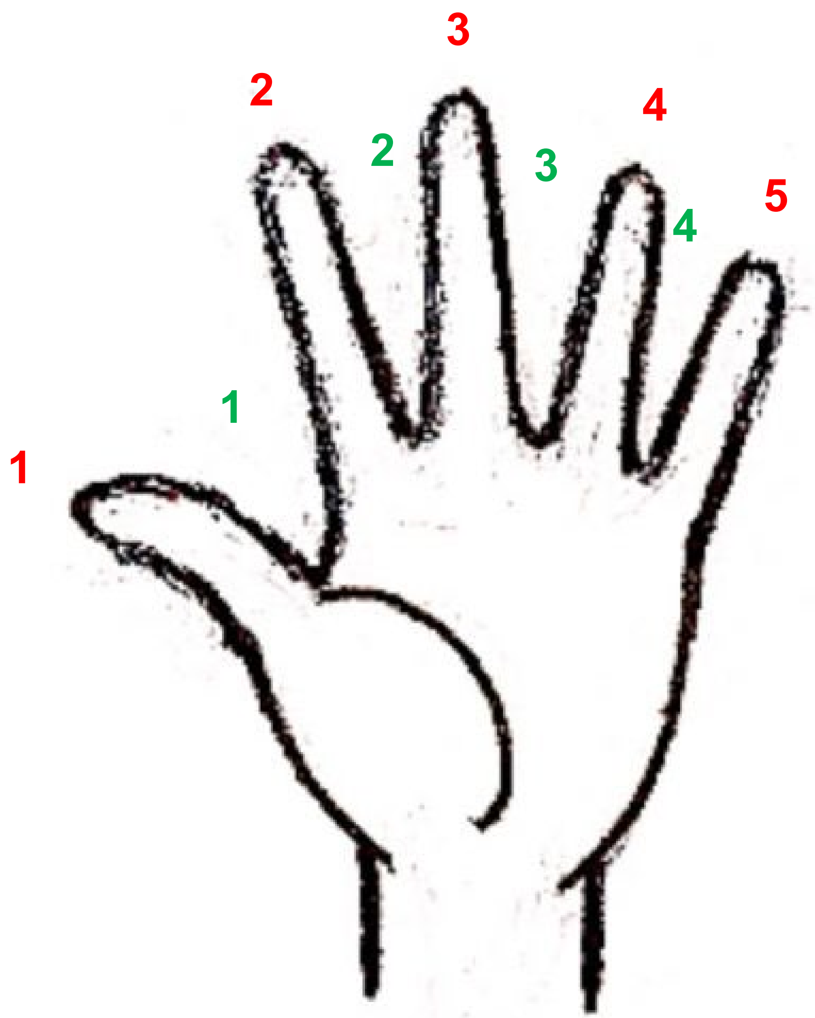
- Stochastic ordering (fit of data to model)
- Monotonicity (ordering or response options)
- No local response dependencies or LID (no significant correlations between items)
- Unidimensionality (one latent construct)
- No differential item functioning or DIF (no sample subgroup effects)

Difficulty Thresholds

Difficulty thresholds are the equal probability points which separate two adjacent response levels in a questionnaire item.

Difficulty Thresholds

Example: A questionnaire with **5 response options** would have **4 thresholds**.

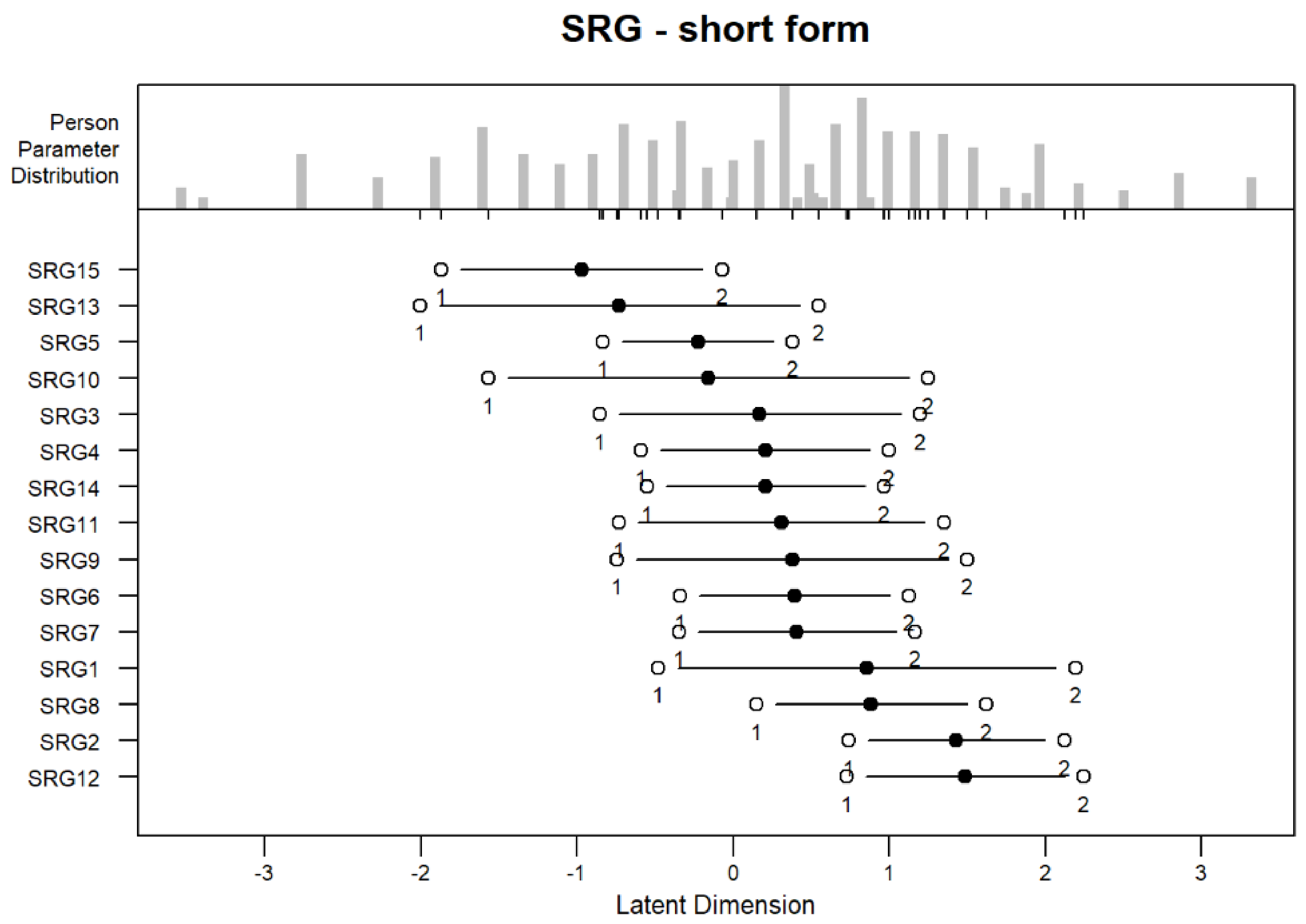


Difficulty Thresholds

The PCM, partial credit model allows non-equidistant thresholds.

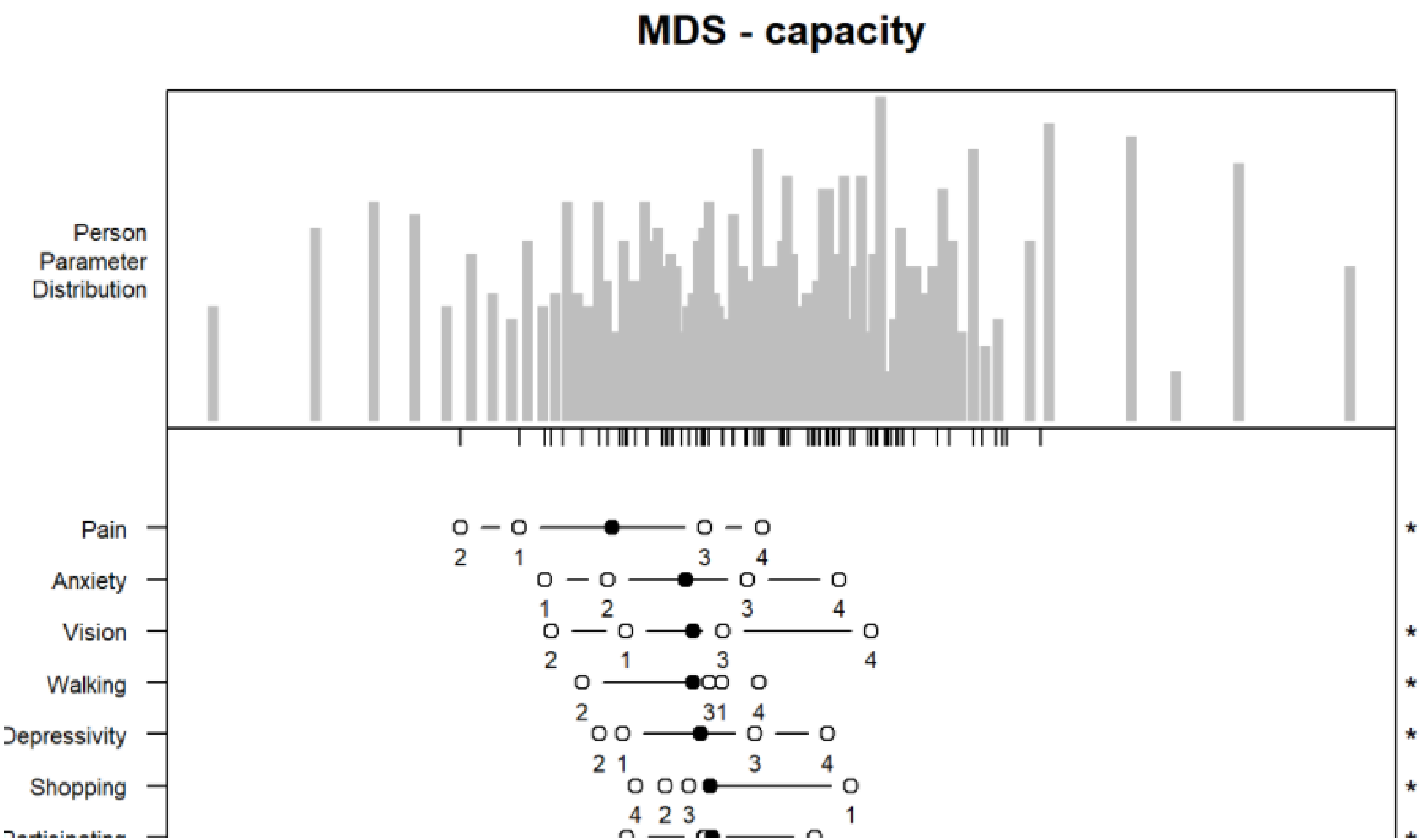
Reversing and disordering of thresholds can happen for example with many response options, with vaguely defined response options, participants with unexpected response behaviors...

Example for Ordered Thresholds:
Person Item Map



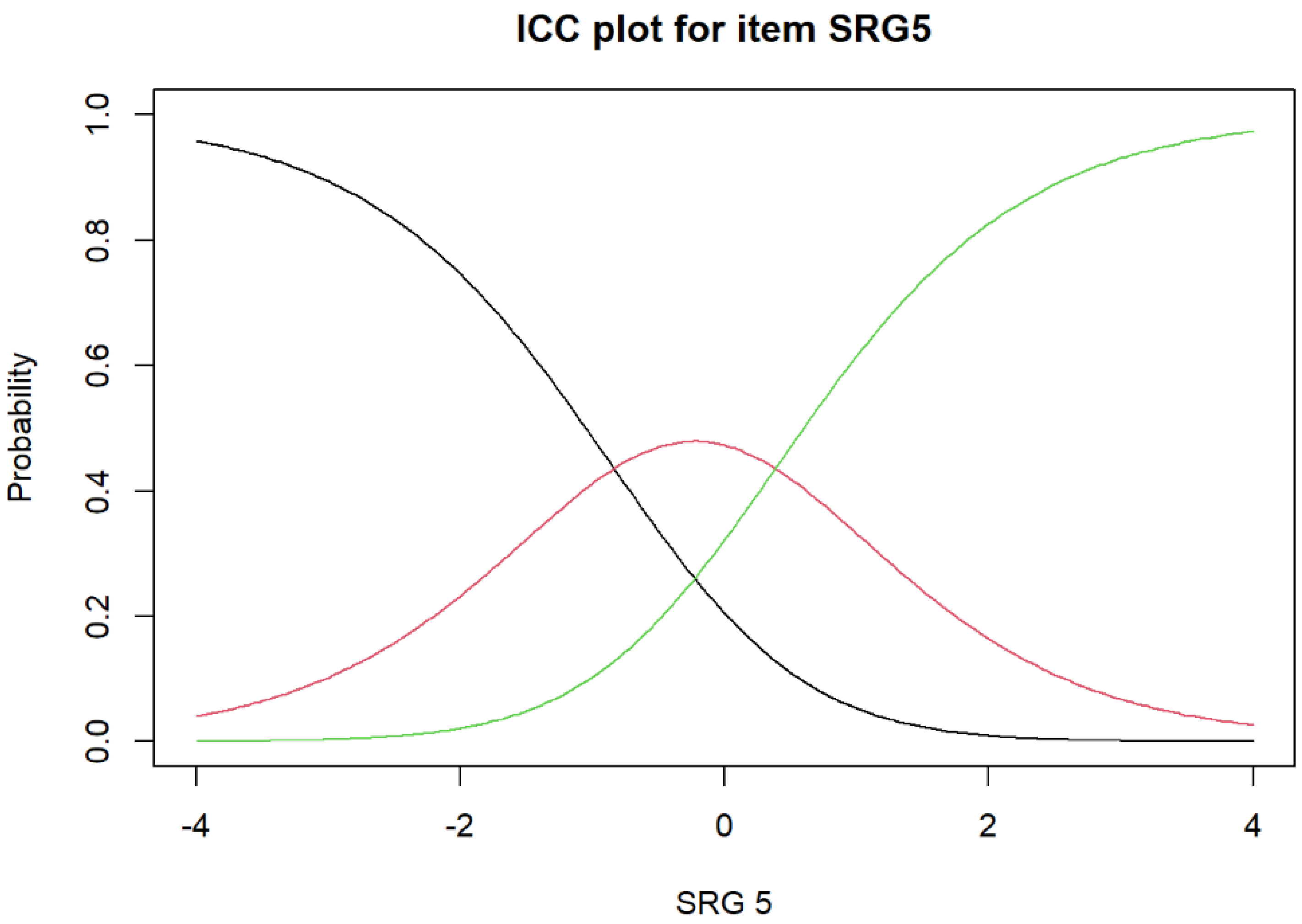
Example for Disordered Thresholds

Person Item Map



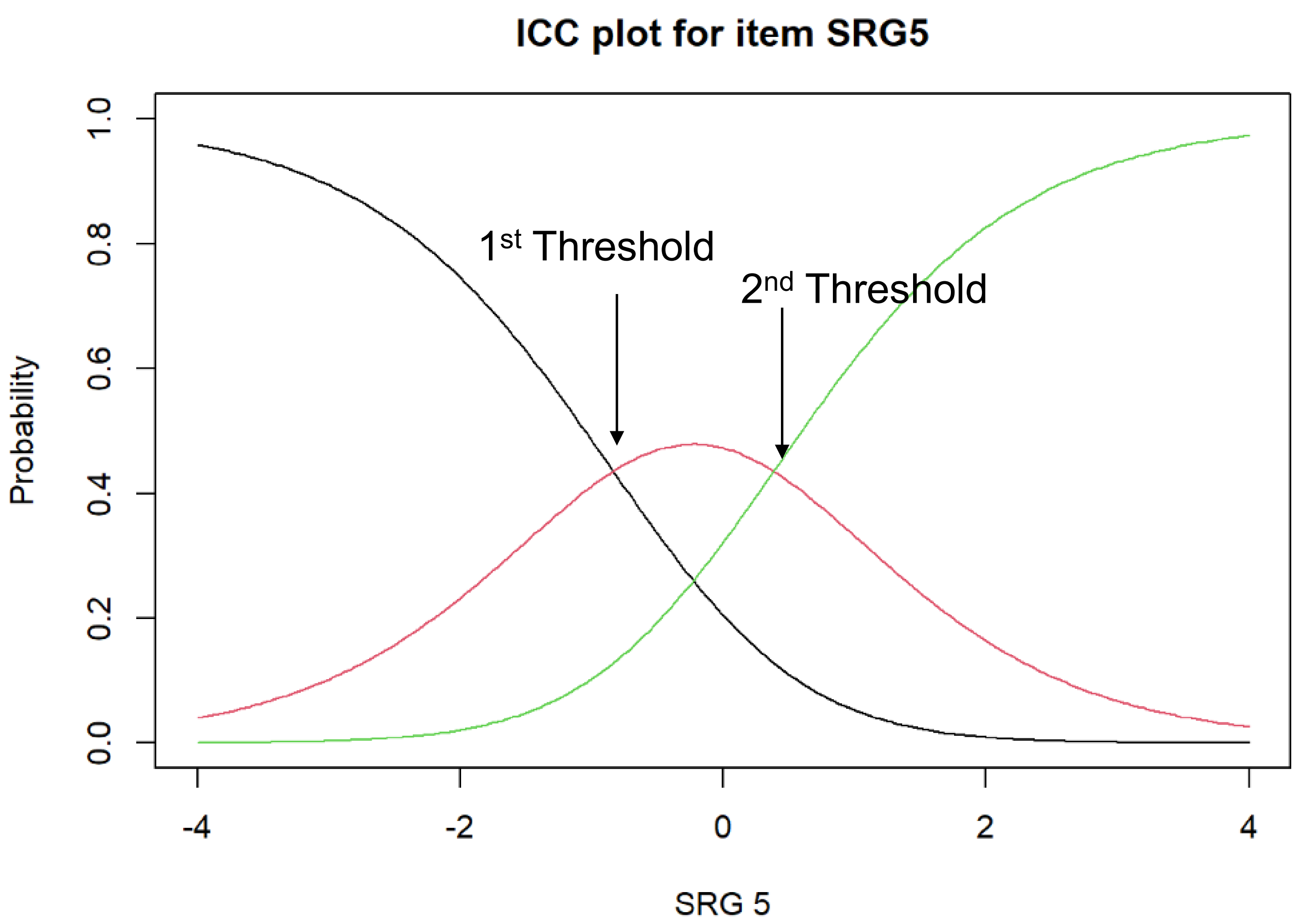
Example for Ordered Thresholds

Item Characteristic Curve



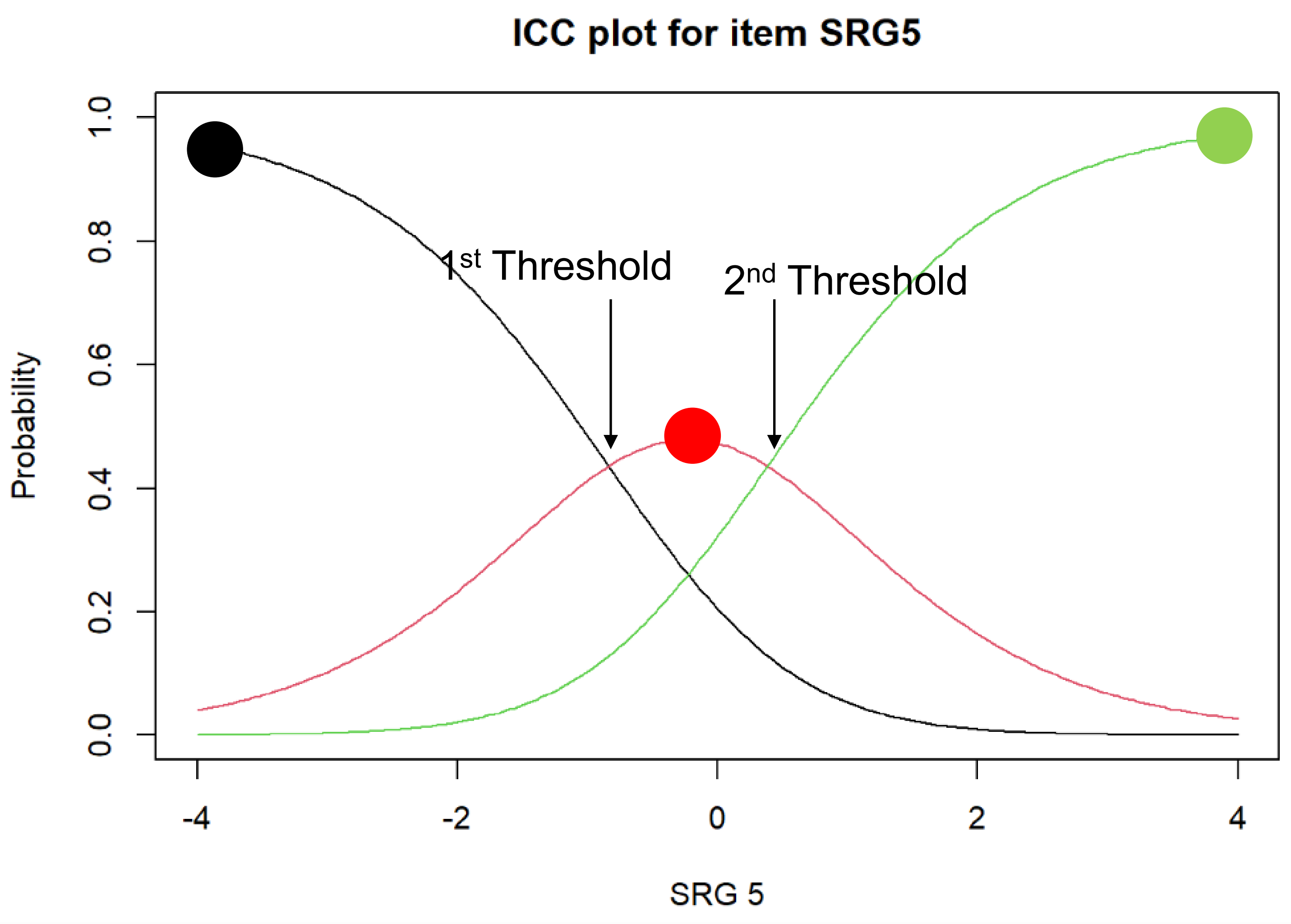
Example for Ordered Thresholds

Item Characteristic Curve



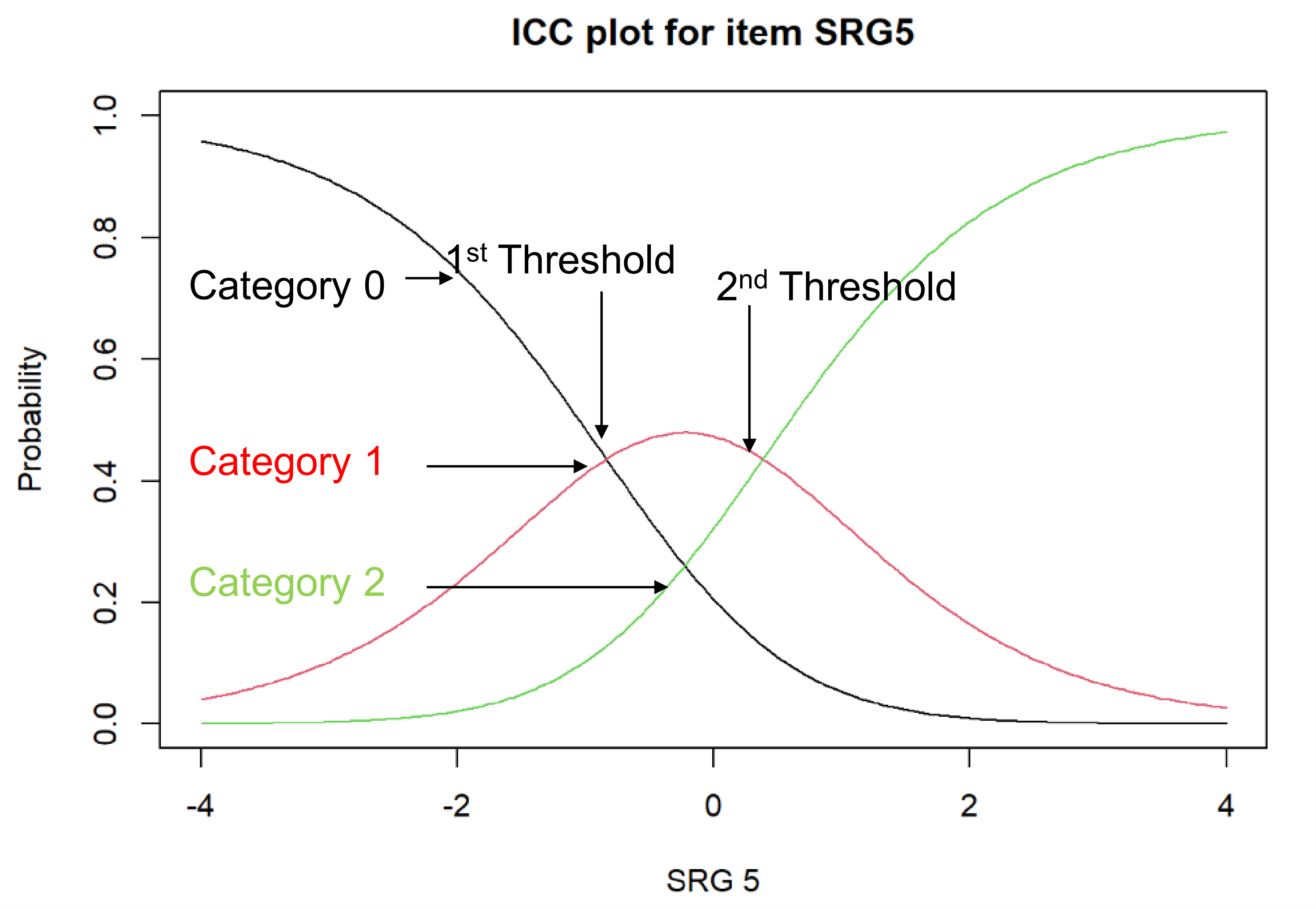
Example for Ordered Thresholds

With many curves it becomes more difficult.
1. Approach: The top of all item category curves are visible

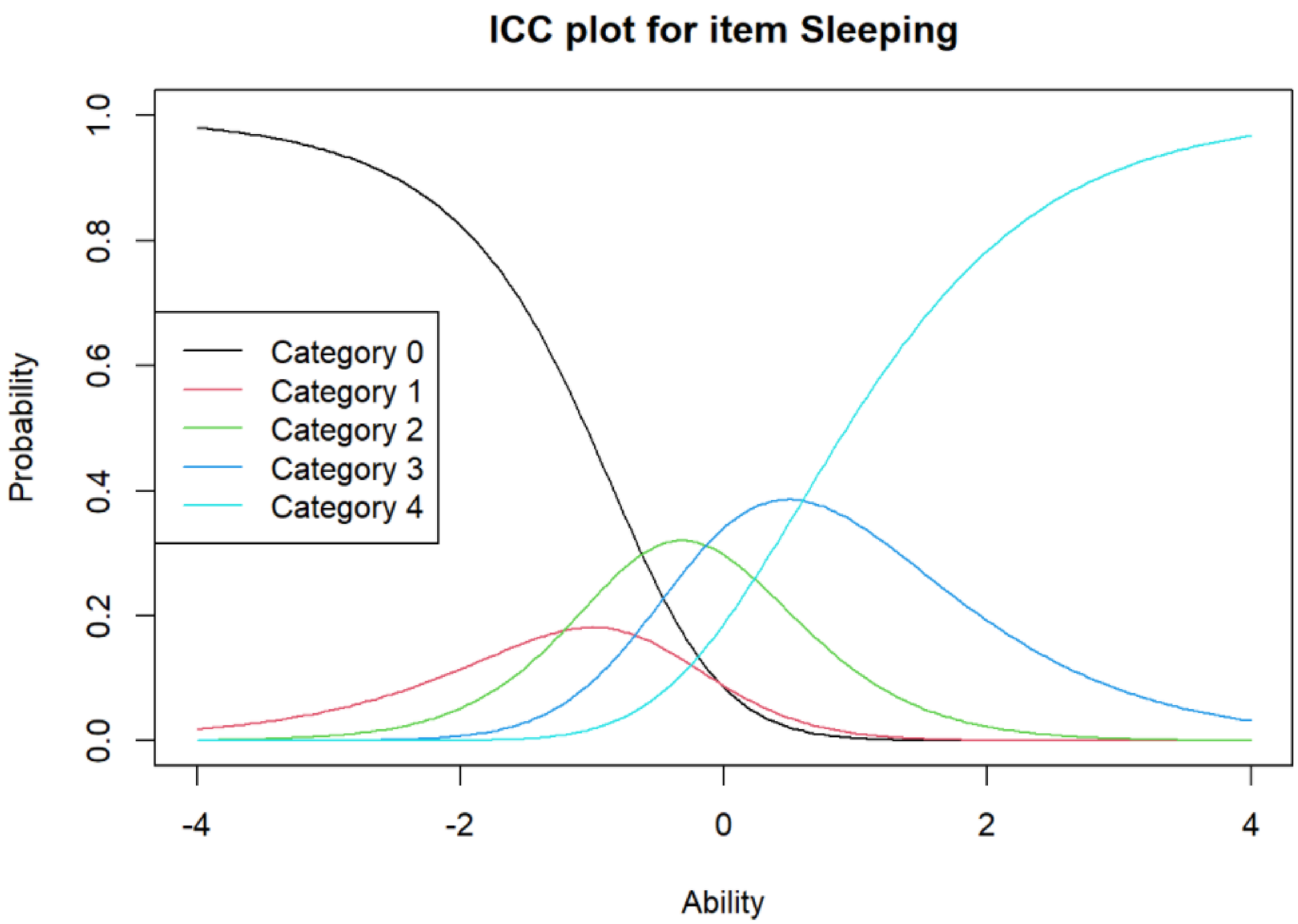


Example for Ordered Thresholds

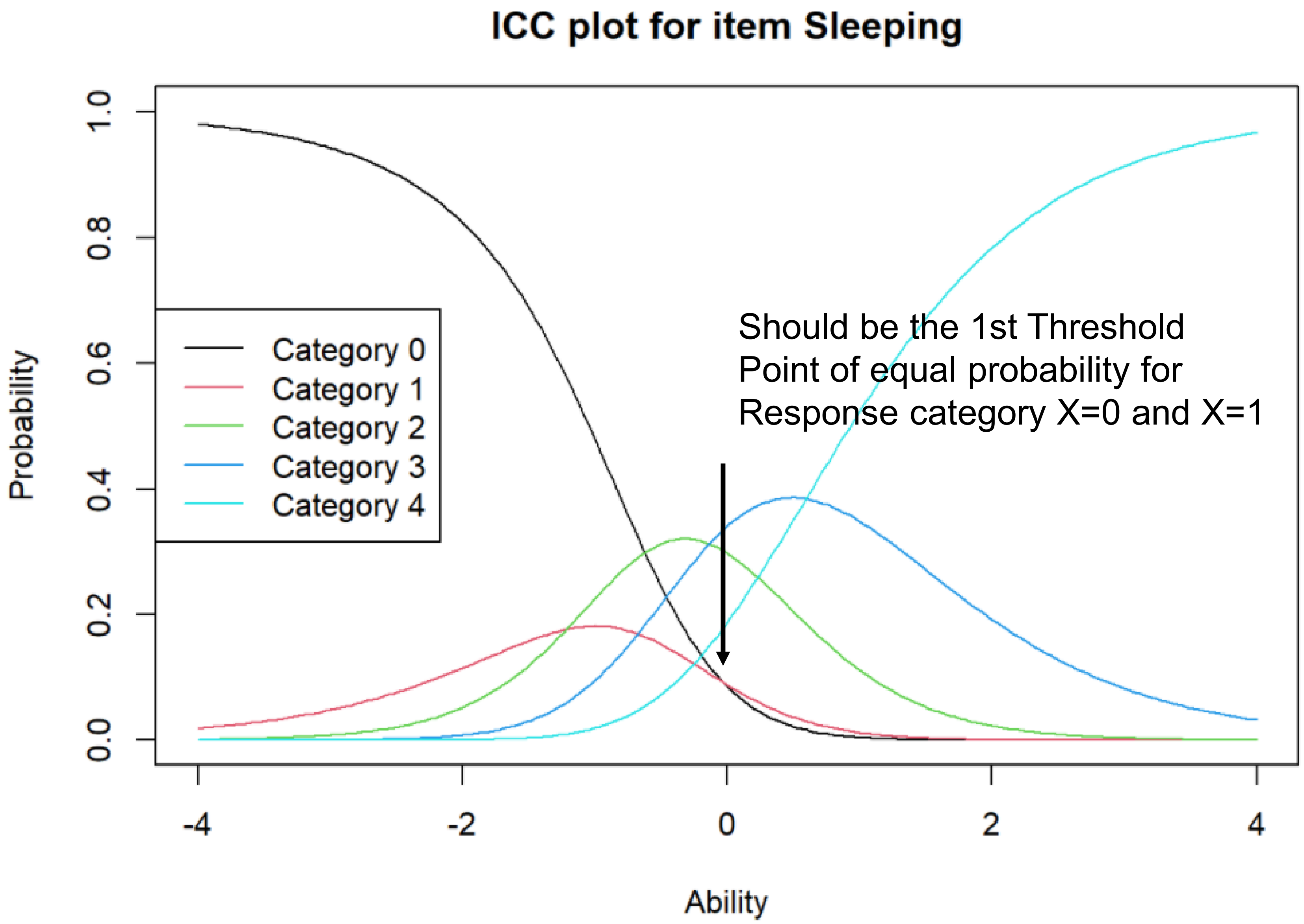
With many curves it becomes more difficult.
2. Approach: The alignment of the intersections on the black line is ordered by response category.



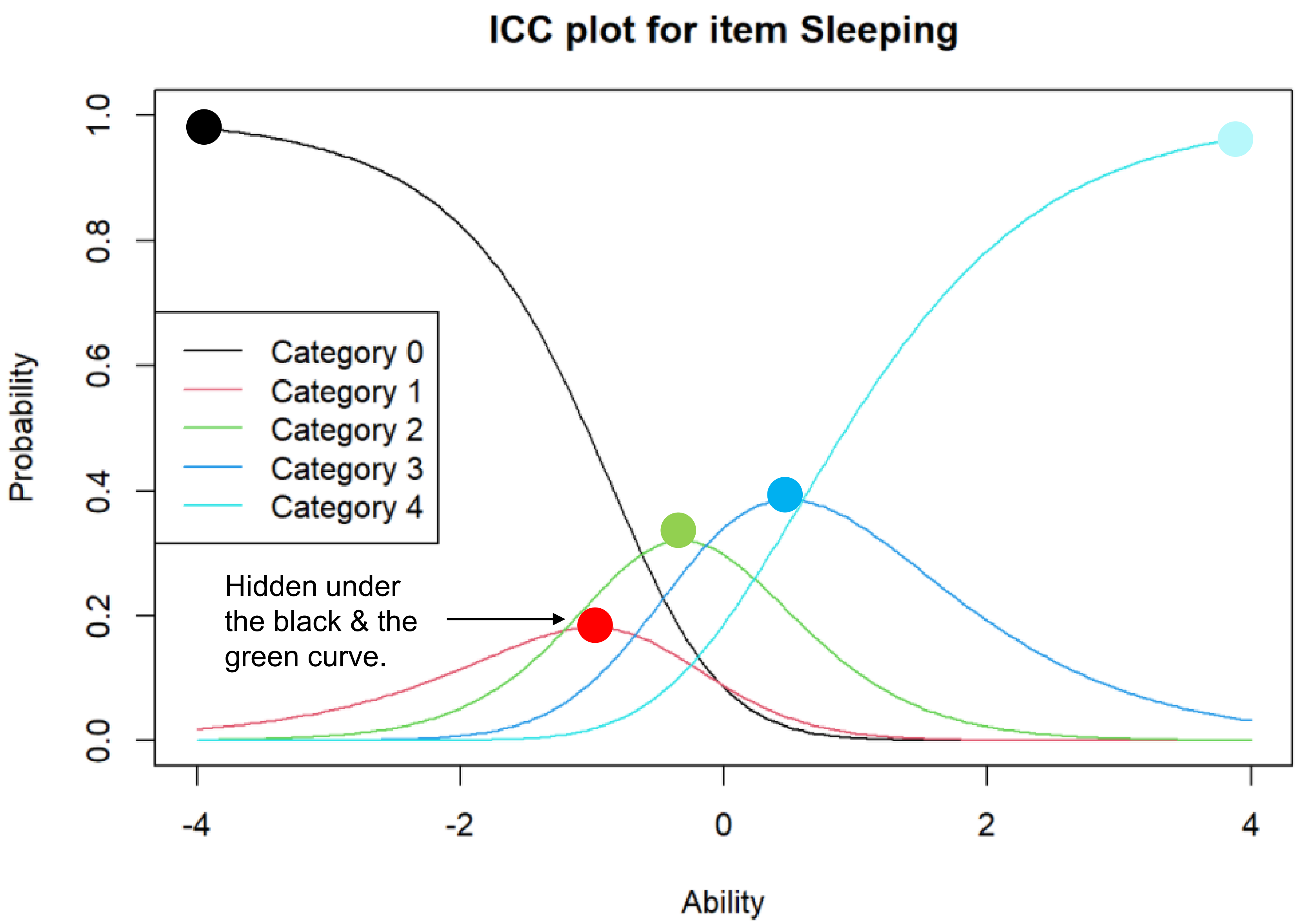
Example for Disordered Thresholds



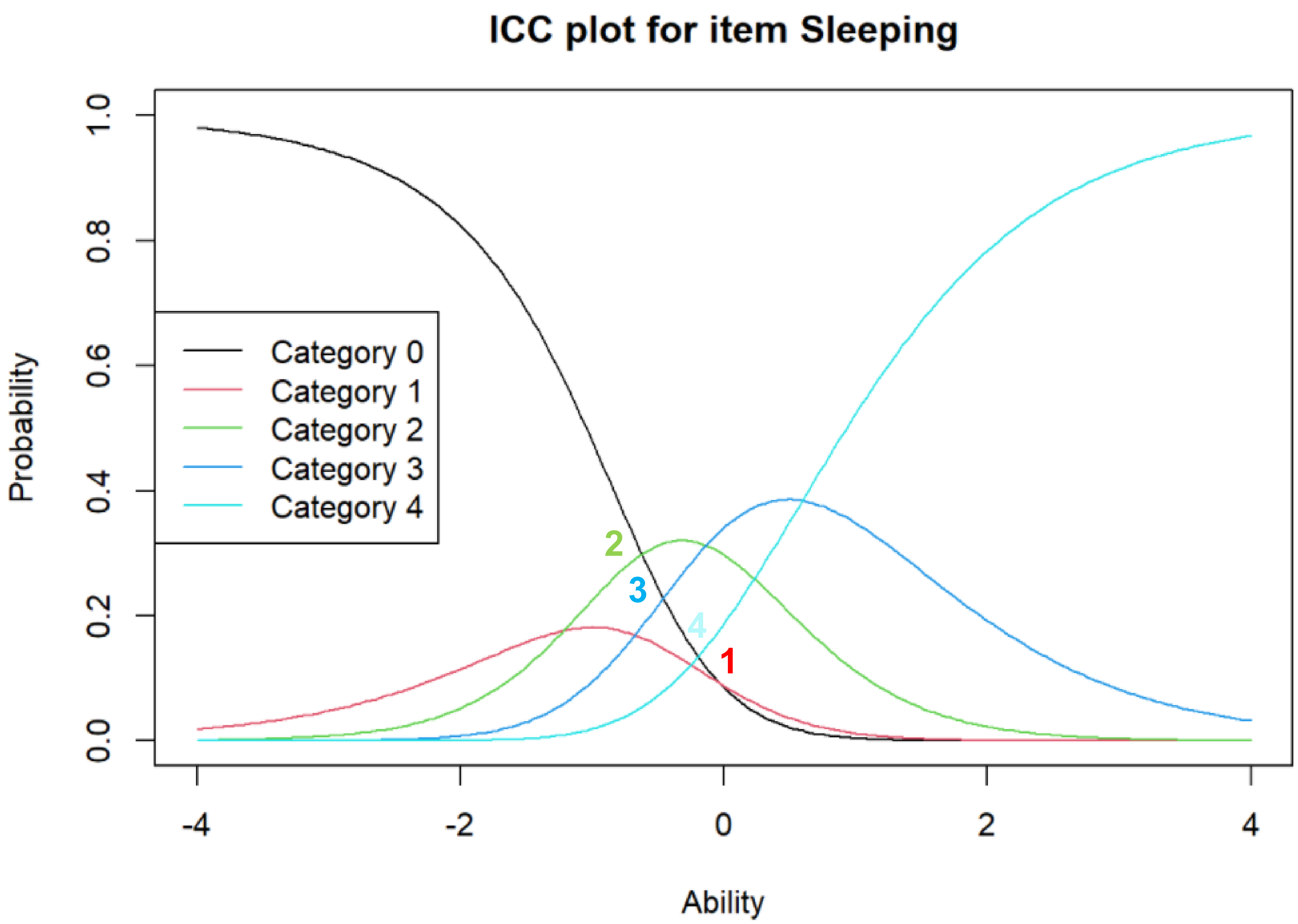
Example for Disordered Thresholds



Example for Disordered Thresholds



Example for Disordered Thresholds



Solving Disordered Thresholds

When the analysis output shows that item thresholds are disordered these can be recoded.

Example: Item original coding: 01234

Output has the first and second threshold that are reversed

Item could be recoded: 00123 or also 01123.

Which option to chose? For example: look at the response frequencies and observe how infit and outfit changes with recoding.

Solving Disordered Thresholds

Not collapse the disordered response options of all items in one step.

Start with the item(s) showing the worst item fit statistics and then proceed stepwise. Sometimes solving disordering in some items improves the ordering of other items.

In some circumstances where disordering affects an entire scale with same or similar disordering across items, a «global» strategy is better. All items are then recoded at once in a same way.

A certain amount of trial and error is sometimes necessary to come to a solution.

Rasch Analysis

A serie of assumptions have to be tested. If the scale ratings comply to these assumptions, the total score is interval-scaled.

- Stochastic ordering (fit of data to model)
- Monotonicity (ordering or response options)
- No local response dependencies or LID (no significant correlations between items)
- Unidimensionality (one latent construct)
- No differential item functioning or DIF (no sample subgroup effects)

Local Item Dependence

Local item dependencies (LID) indicate that pairs of items are associated or correlated above a certain cut-off.

LID introduces bias in the estimation of the reliability of the metric.

Residual Correlation (Q_3)

Strength of item association is computed using the correlation matrix of the standardized residuals.

Items are said locally dependent when correlating positively above a certain cut-off.

The cut-off is typically set at 0.2 or 0.3

$$corr(X, Y) = \frac{cov(XY)}{\sigma_x \sigma_y}$$

	I1	**I2**	**I3**
I1	1	0.03	0.4
I2	0.03	1	-0.3
I3	0.4	-0.3	1

Example of a correlation matrix

Residual Correlation (Q_3)
Cut-off

The cut-off for an acceptable item residual correlation is typically set at 0.2 or 0.3.

Recent simulation studies have suggested another, more reliable but more conservative, approach to detect LID:

$$Q_3^{\star} = Q_{3,max} - \bar{Q}_3 > 0.2$$

The cut-off corresponds to the mean residual correlation + 0.2. No residual correlation should be above this cut-off.

Residual Correlation (Q_3)
Visual inspection

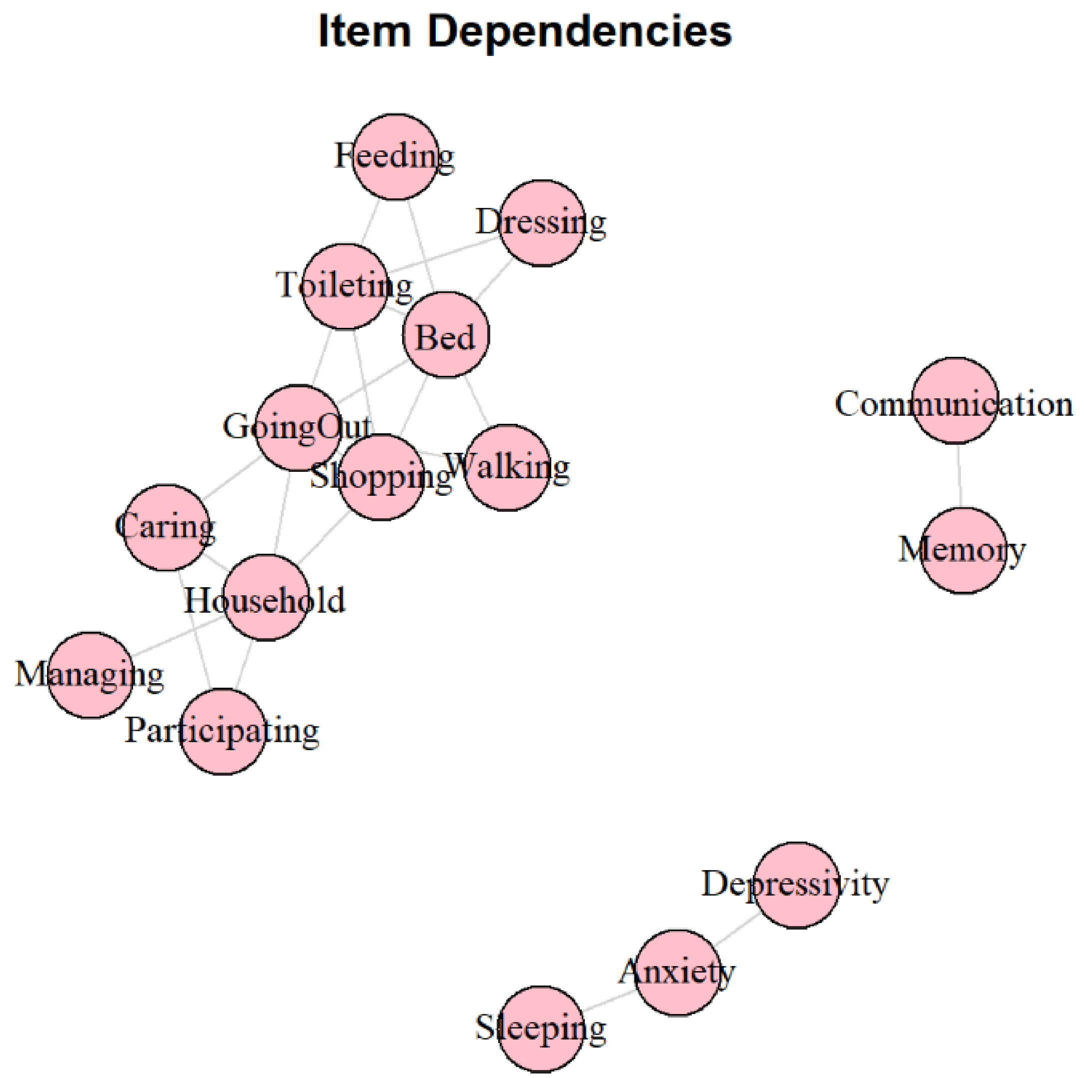
One approach to detect the pairwise dependencies, is to search through the correlation matrix.

With large scales, the inspection of the residual matrix can become tedious.

Another approach is to **visualize the dependencies** with a graphical model.

The graphical model has the advantage to show association patterns, beyond the pairwise correlations.

Residual Correlation (Q_3) Visual inspection



Residual Correlation (Q_3)
Solving Dependencies: Testlets

LID above cut-off inflates the reliability estimates. LID is strongly related to multidimensionality.
To solve the items dependencies:

1) Very redundant items could be deleted.
2) Creation of Testlets.

Advise to not delete any scale items and to create testlets.
Testlets consist of the sum score of the dependent items.

The individual items are removed and enter the analysis as one aggregated testlet.

Rasch Analysis

A serie of assumptions have to be tested. If the scale ratings comply to these assumptions, the total score is interval-scaled.

- Stochastic ordering (fit of data to model)
- Monotonicity (ordering or response options)
- No local response dependencies or LID (no significant correlations between items)
- Unidimensionality (one latent construct)
- No differential item functioning or DIF (no sample subgroup effects)

Multidimensionality

The Rasch model assumes that a questionnaire measures only one single latent trait or construct.

In presence of multidimensionality, the scale measures different aspects of a construct and single interval scaled sum score is not meaningful anymore.

Standardised Residuals

The analysis for multidimensionality searches the standardised residuals for patterns indicating items loading strongly on different dimensions.

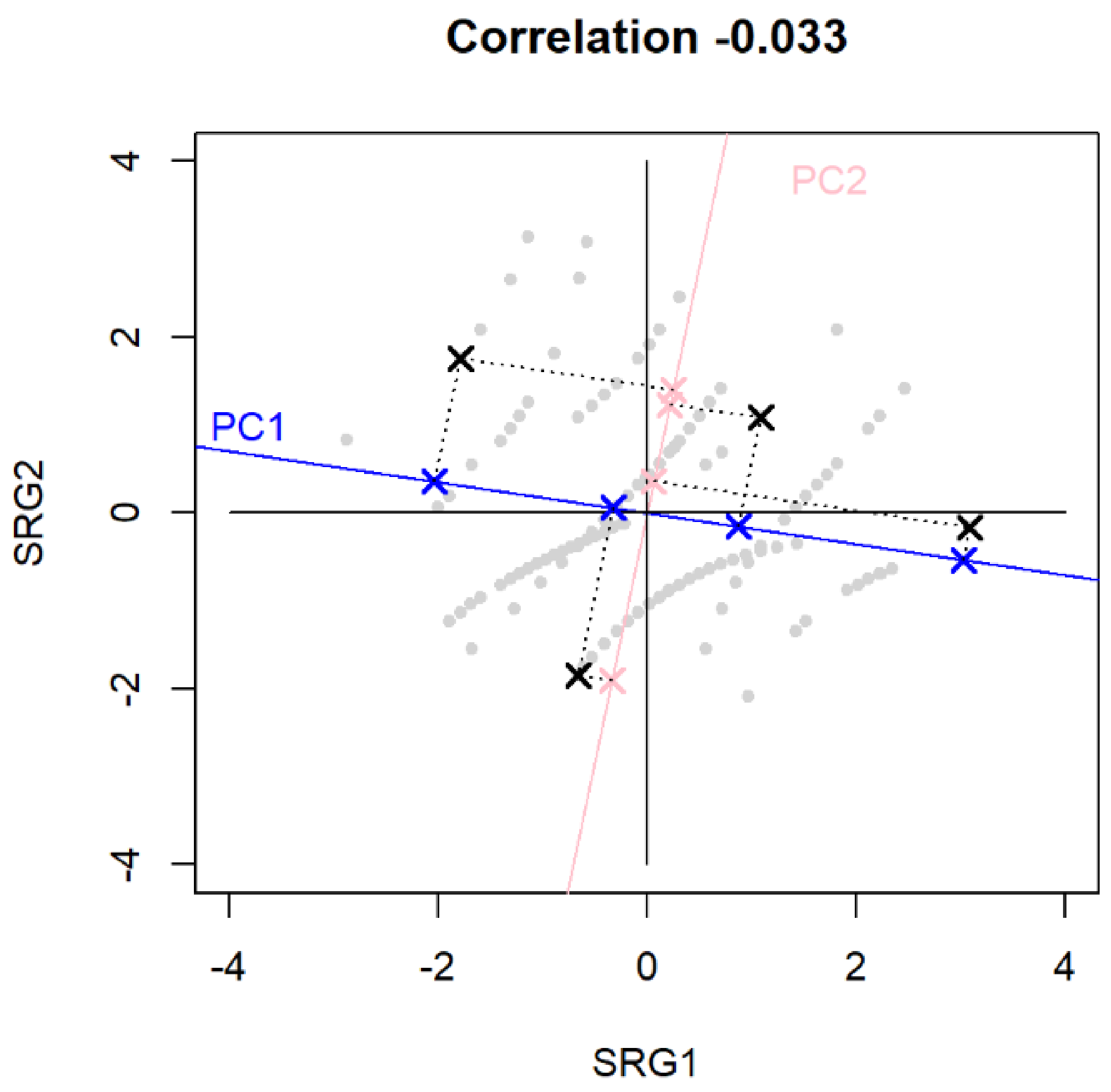
One method to analysis the standardised residuals is called principal component analysis (PCA).

Principal Component Analysis (PCA)

- is a dimensionality reduction technique
- allow to identify clusters of similar variables.
- needs no distributional assumptions.
- is an exploratory method bases on singular value decomposition (SVC) or eigendecomposition.

Central idea: reduce the dimensionality of a dataset, while preserving as much ‘variability’ (i.e. statistical information) as possible, i.e. through maximizing the variance in each dimension.

Principal Components

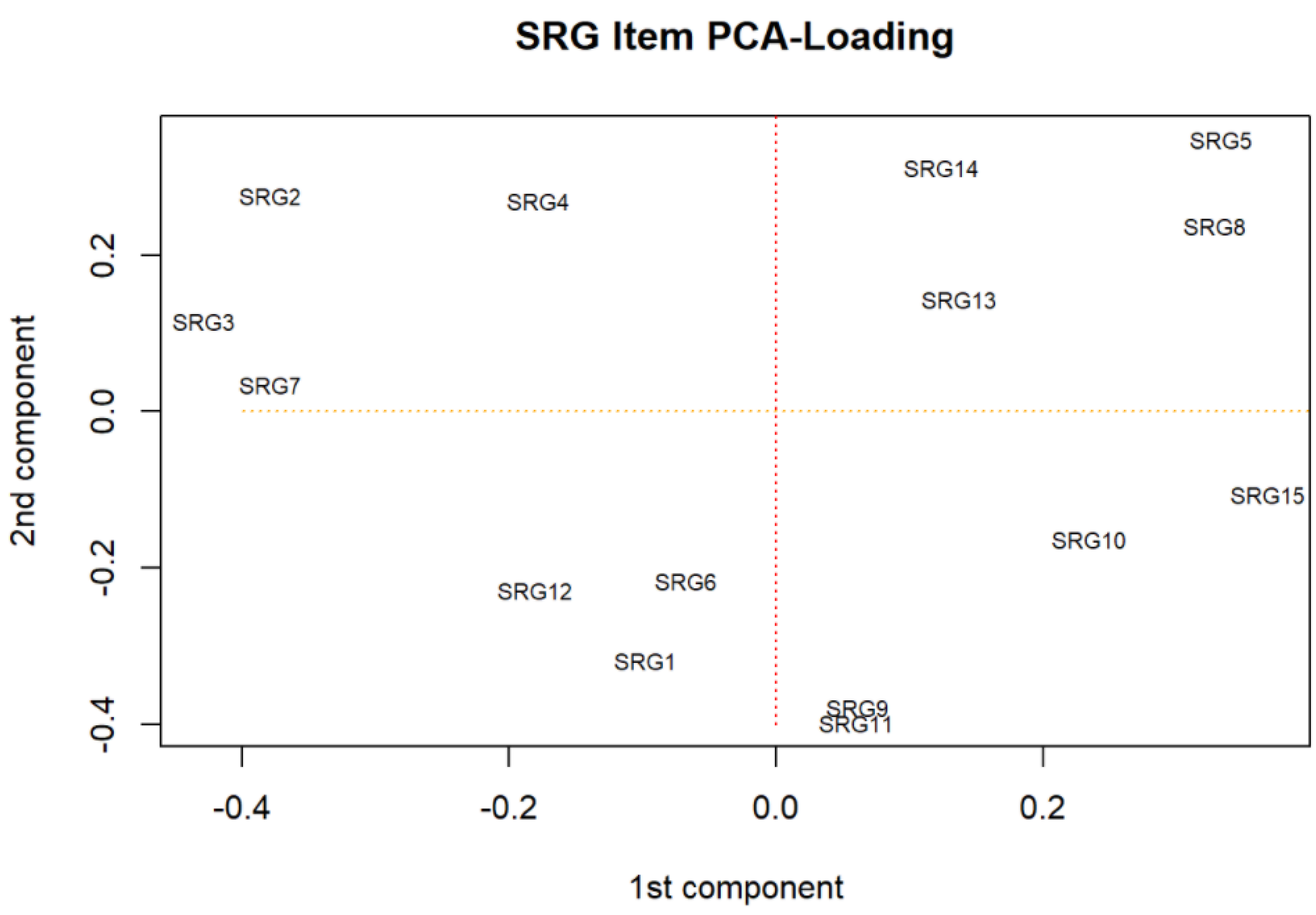


Component loading Matrix (or eigenvector)

	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9	PC10
SRG1	-0.098	-0.319	-0.235	-0.349	0.174	0.196	0.348	-0.212	0.189	0.487
SRG2	-0.379	0.276	-0.215	0.095	0.343	-0.043	0.022	-0.198	0.063	0.098
SRG3	-0.429	0.116	0.336	-0.072	-0.107	0.283	-0.068	0.213	0.160	0.121
SRG4	-0.178	0.269	-0.322	-0.066	-0.322	0.107	0.284	0.556	0.103	-0.255
SRG5	0.333	0.348	-0.091	0.175	-0.174	-0.081	-0.281	0.122	-0.258	0.486
SRG6	-0.068	-0.216	0.403	0.375	0.048	0.054	0.409	-0.089	-0.211	-0.223
SRG7	-0.379	0.035	0.013	0.039	-0.420	-0.128	-0.264	-0.503	-0.195	0.030
SRG8	0.328	0.238	0.131	0.317	0.094	0.210	0.374	-0.129	-0.126	0.076
SRG9	0.061	-0.379	-0.205	0.198	-0.138	0.454	-0.371	-0.091	0.057	-0.328
SRG10	0.234	-0.164	-0.160	0.357	-0.169	-0.323	0.026	-0.065	0.690	0.043
SRG11	0.060	-0.398	-0.363	-0.135	-0.068	-0.278	0.149	0.146	-0.511	-0.046
SRG12	-0.181	-0.229	0.237	0.125	0.428	-0.408	-0.280	0.379	0.015	0.044
SRG13	0.137	0.143	0.279	-0.451	-0.223	-0.413	0.158	-0.214	0.126	-0.276
SRG14	0.124	0.312	-0.264	-0.124	0.481	0.032	-0.146	-0.195	-0.001	-0.436
SRG15	0.368	-0.106	0.302	-0.409	0.042	0.271	-0.228	0.092	0.054	0.036

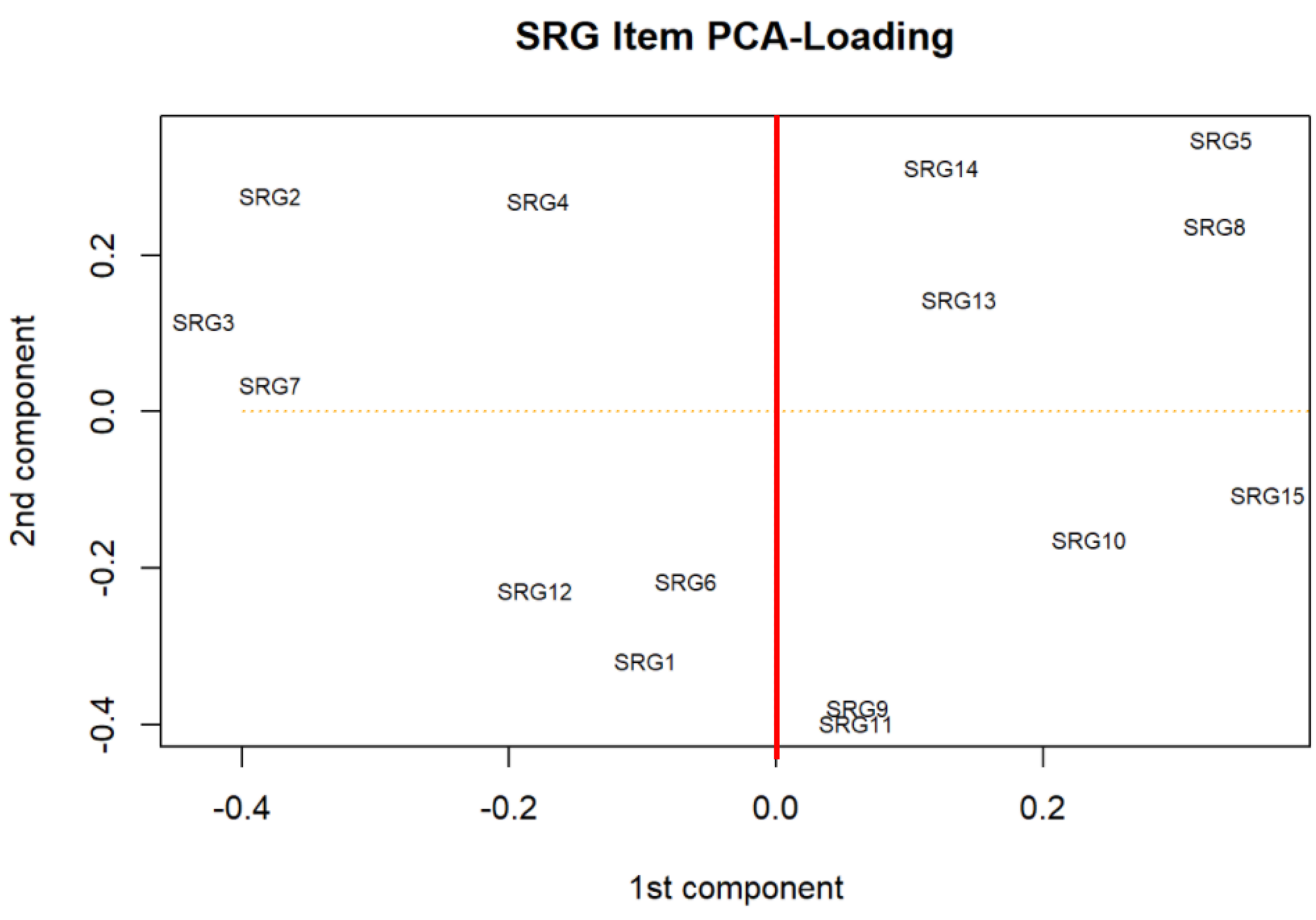
The residual matrix is factorized into several component matrices, including eigenvectors or component loading matrix. Component loading matrix has as many columns and rows as items in the scale (here column 11 to 15 are not shown.) PC1 explains most of the variability in the residuals, it is the most important. PC2 explains what is left unexplained from PC1 etc… etc…

PCA: Component Loading



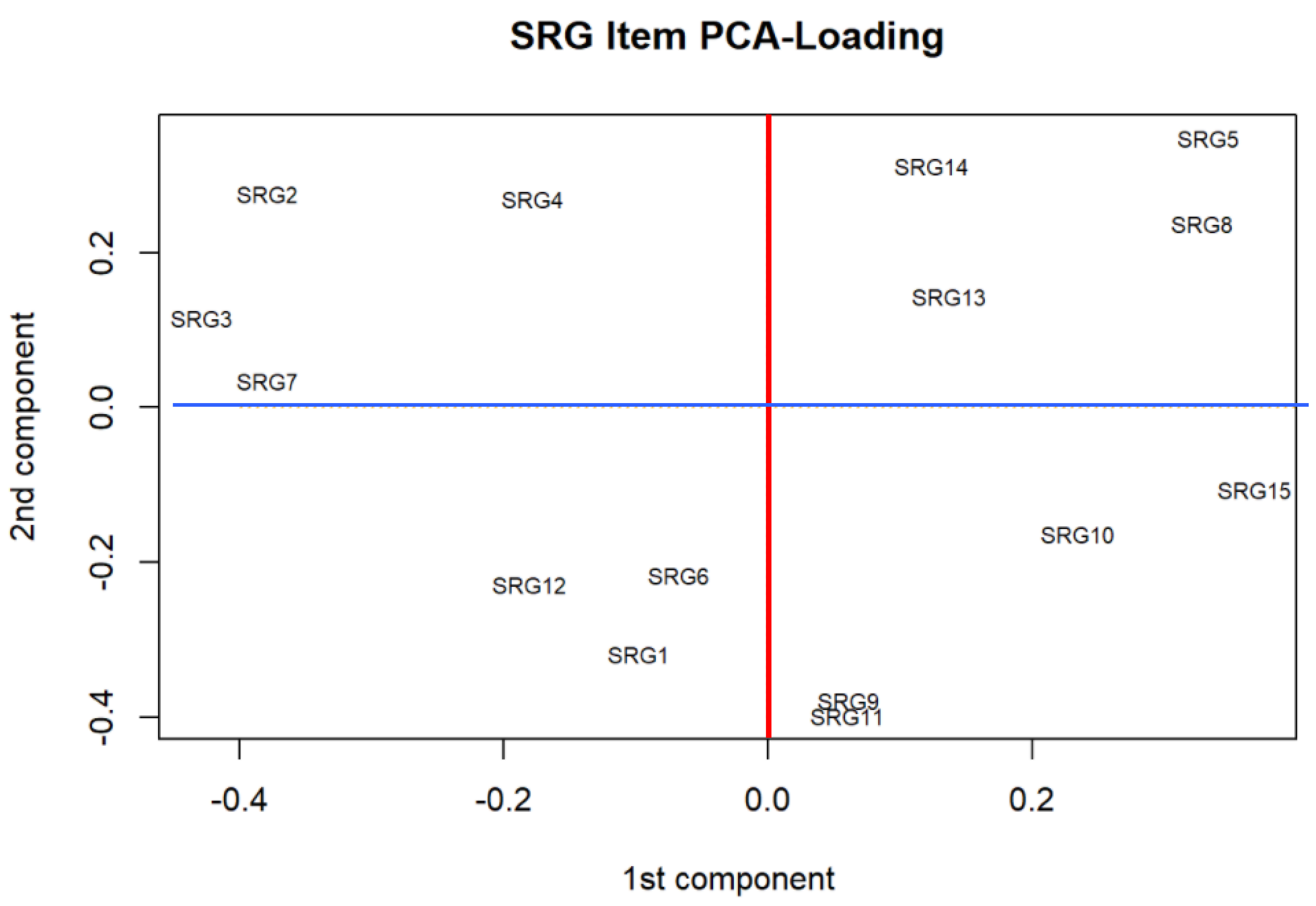
The first 2 columns of the component loading matrix provides the x and y coordinates for the plot above.

PCA: Component Loading



The opposition on the x-axis is the most important.

PCA: Component Loading



The opposition on the x-axis is the most important.

Eigenvalues

- The component loadings do not allow to determine if the display is unidimensional or indicative of multidimensionality.

- The eigenvalue vector allows to determine if a set of items is unidimensional or multidimensional.

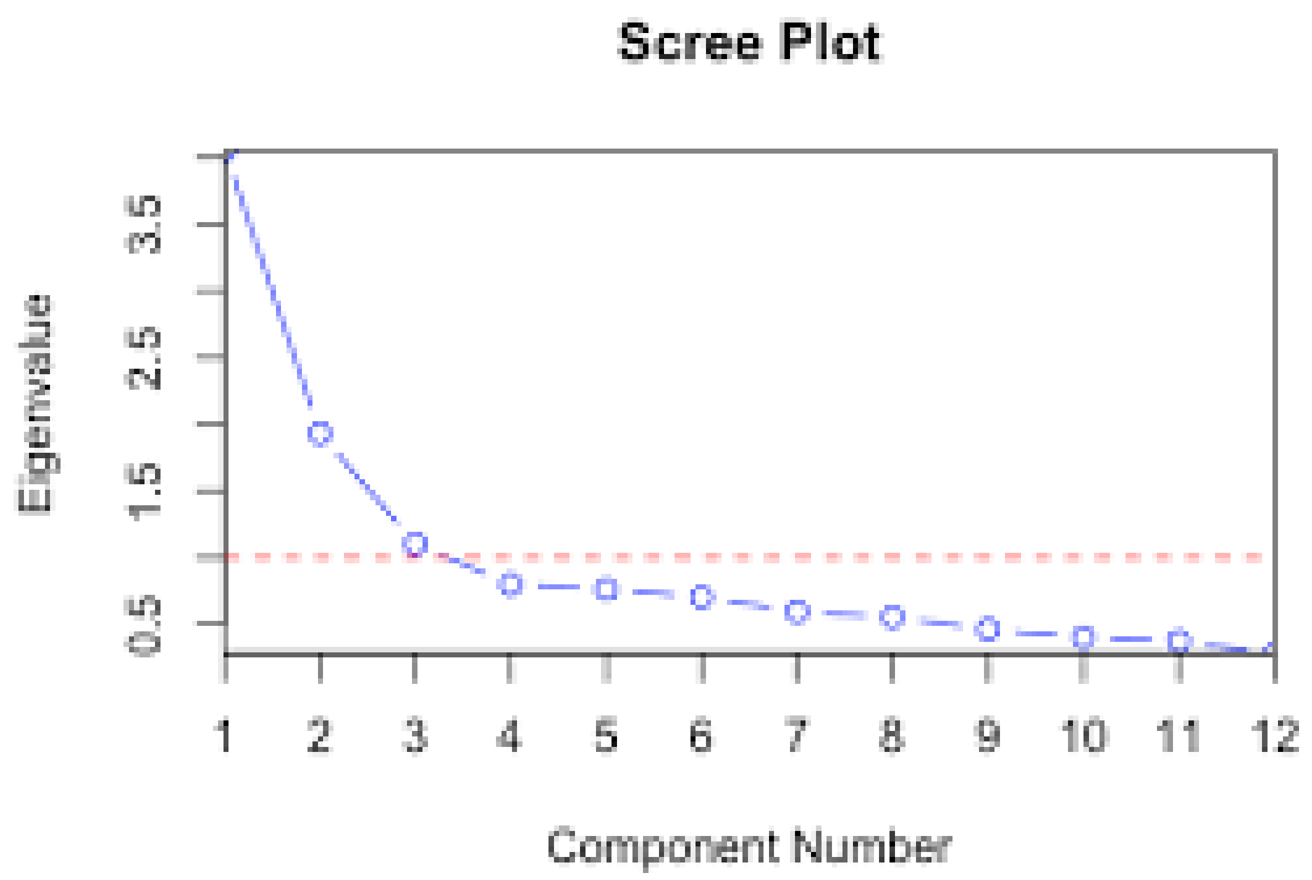
- Diverse rules are available to interpret the eigenvalue vector.

    - the first eigenvalue should not be too large, at least < 2
    - Analysis of a screeplot to determine the number of dimensions – number of components left of the elbow

Eigenvalues

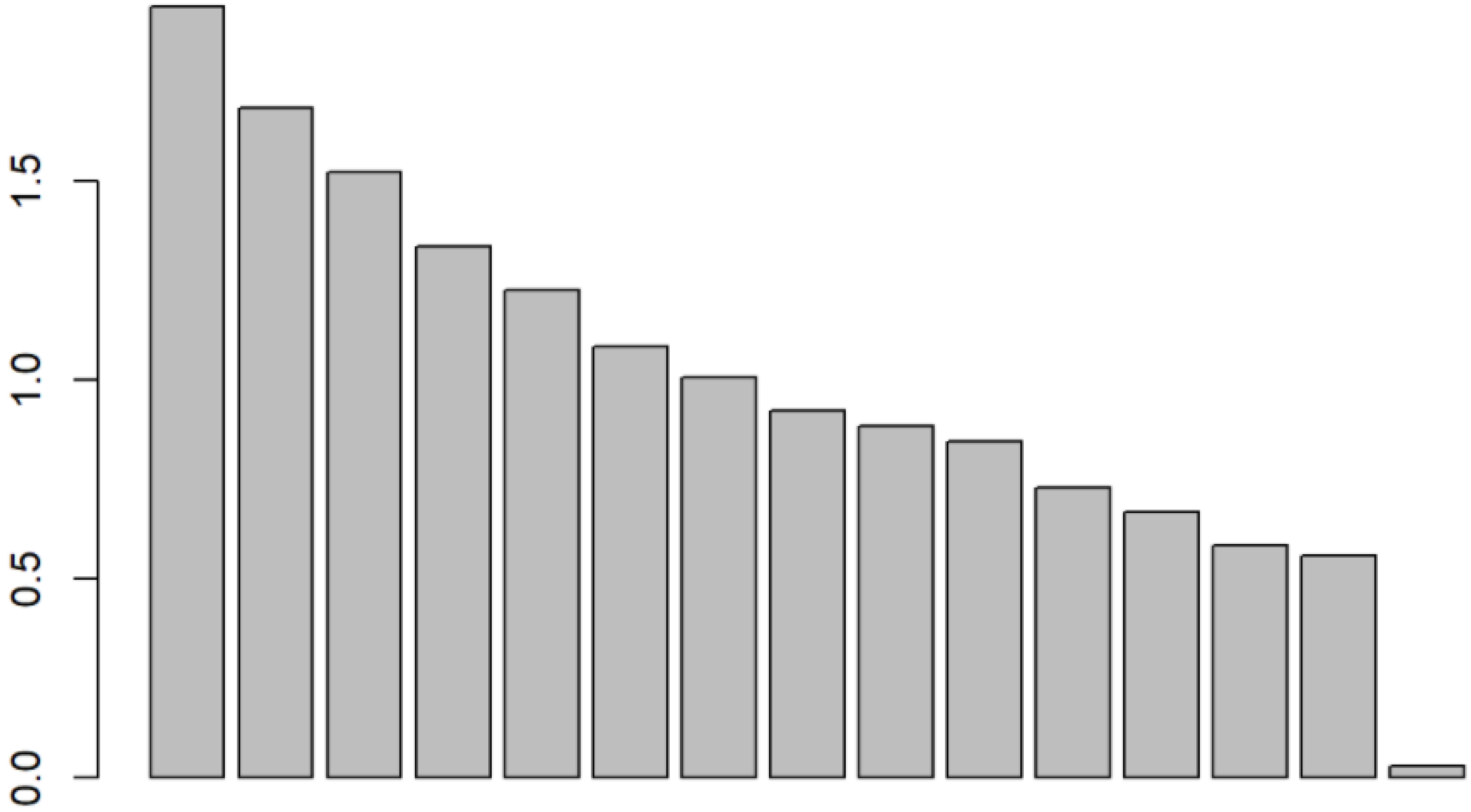
	Eigen.Value.srg	Perc.Eigen.srg	Cum.Perc.Eigen.srg
[1,]	1.93862672	12.9241781	12.92418
[2,]	1.68288091	11.2192061	24.14338
[3,]	1.52202583	10.1468389	34.29022
[4,]	1.33477400	8.8984933	43.18872
[5,]	1.22401820	8.1601213	51.34884
[6,]	1.08352636	7.2235090	58.57235
[7,]	1.00627240	6.7084827	65.28083
[8,]	0.92081890	6.1387926	71.41962
[9,]	0.88194488	5.8796325	77.29925
[10,]	0.84505677	5.6337118	82.93297
[11,]	0.72765528	4.8510352	87.78400
[12,]	0.66557609	4.4371739	92.22118
[13,]	0.58298155	3.8865437	96.10772
[14,]	0.55707040	3.7138026	99.82152
[15,]	0.02677172	0.1784782	100.00000

Eigenvalues and Screeplot



To determine the number of dimensions a rule is to determine where the elbow is…
This figure supports more than one dimension, shows about 2 to 3 dimensions left of the line break..

Eigenvalues and Screeplot



To determine the number of dimensions a rule is to determine where the elbow is…
This figure based on SRG does not indicate any strong change in direction.

Rasch Analysis

A serie of assumptions have to be tested. If the scale ratings comply to these assumptions, the total score is interval-scaled.

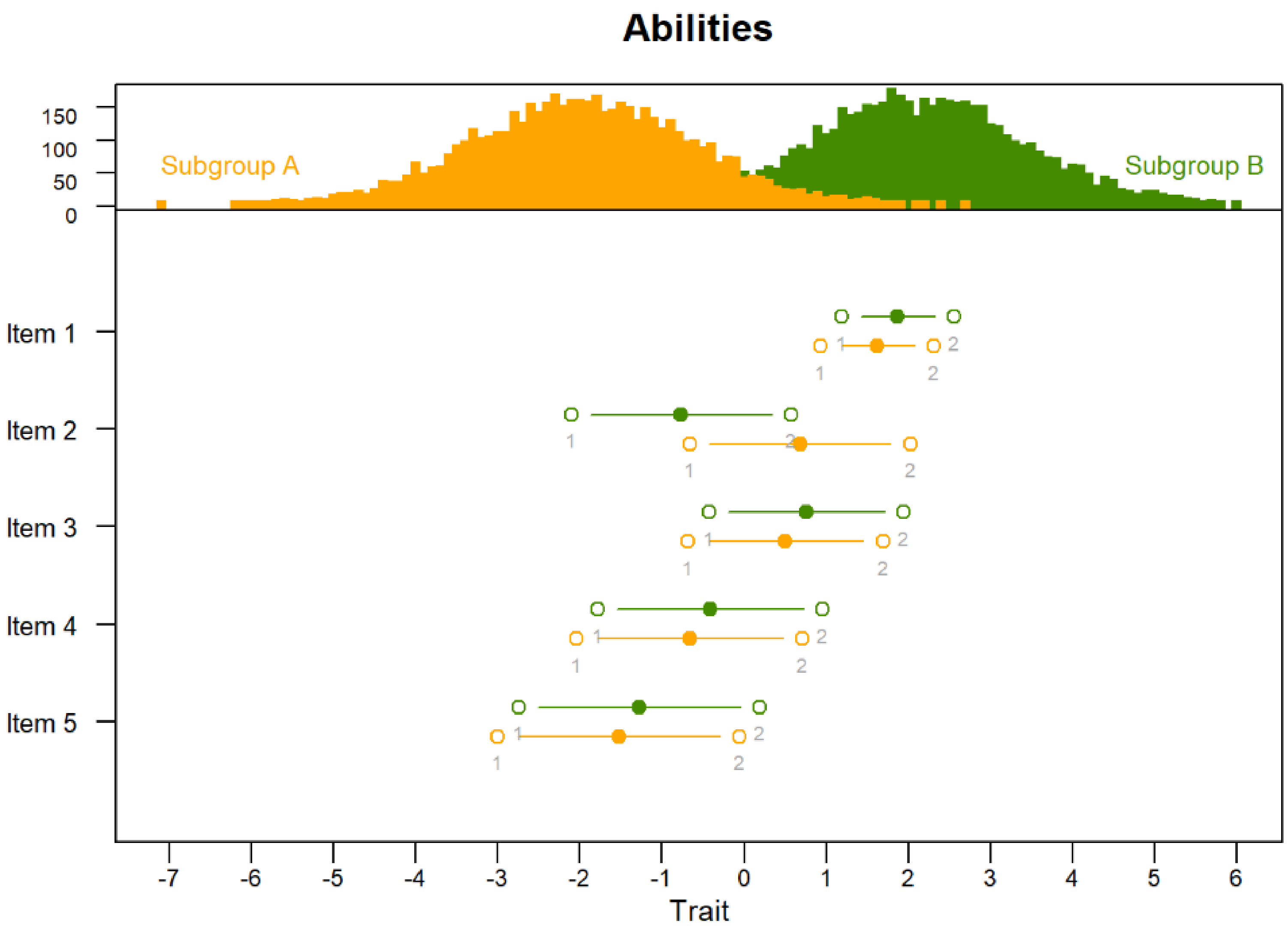
- Stochastic ordering (fit of data to model)
- Monotonicity (ordering or response options)
- No local response dependencies or LID (no significant correlations between items)
- Unidimensionality (one latent construct)
- No differential item functioning or DIF (no sample subgroup effects)

Differential Item Functioning

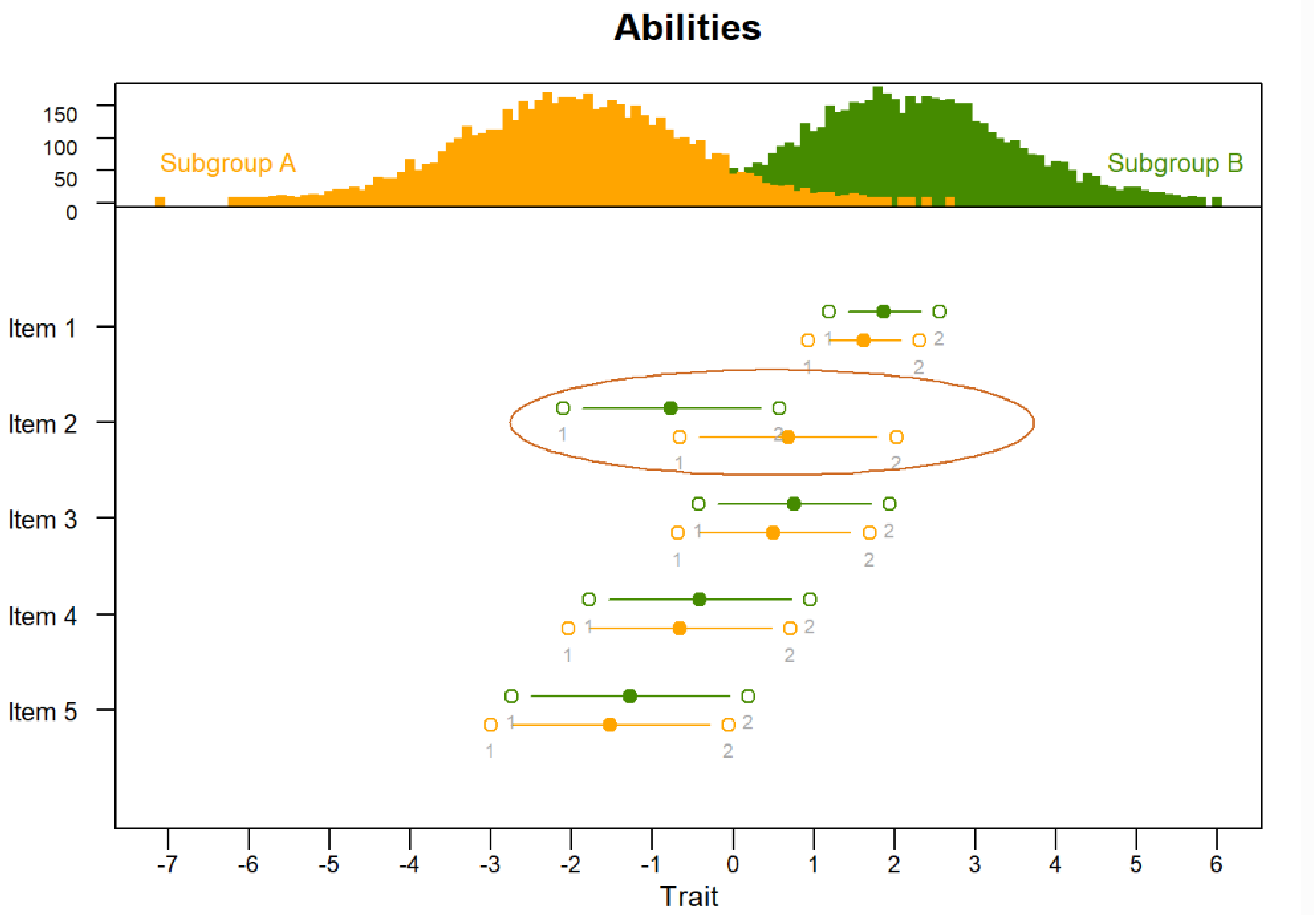
The Rasch model assumes the construct measured is valid across subgroups.

Differential item functioning tests if items are invariant across sample subgroups.

Differential Item Functioning

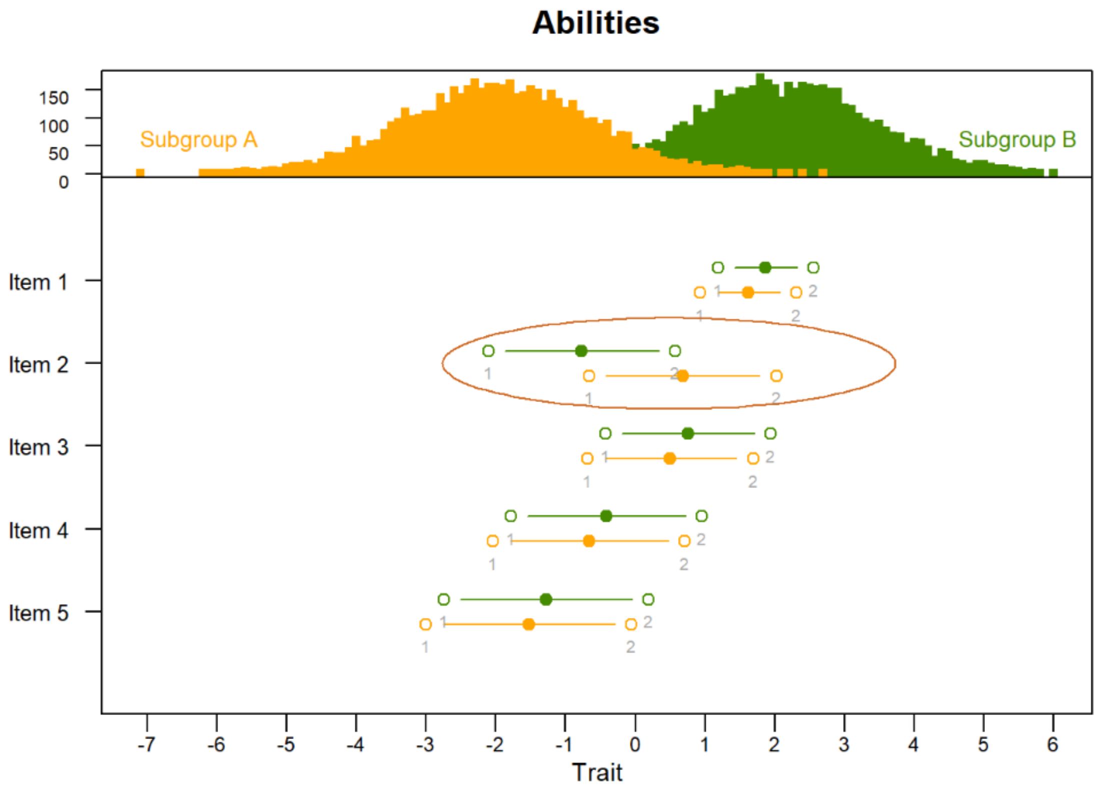


Differential Item Functioning



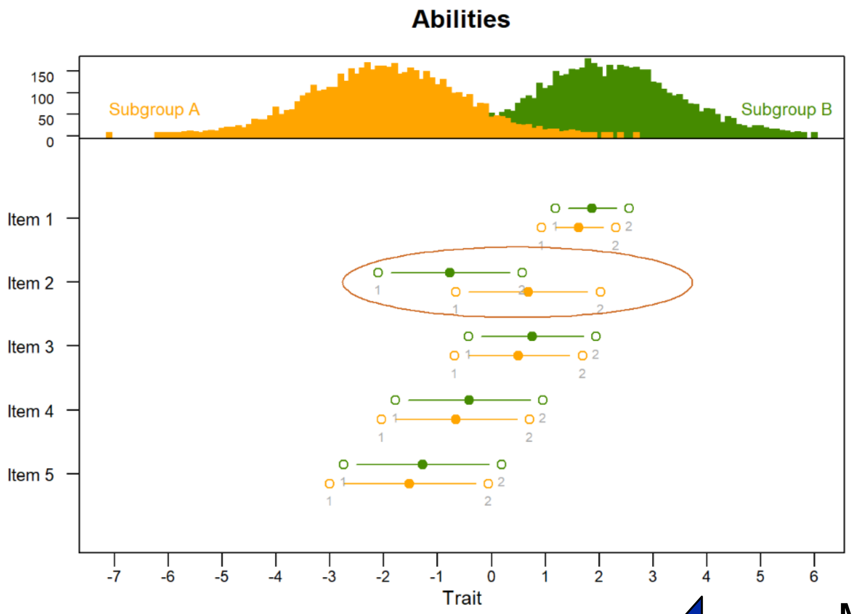
2) one item, the Item 2, is in different locations relative to the other items as a function of the subgroup.

Differential Item Functioning?



1) The ability of Subgroup A is lower than the ability of Subgroup B.
2) The difficulty of the item is similar for almost all items.
3) For a same level of ability, the difficulty of Item 2 differs across the Subgroup A and Subgroup B.

Differential Item Functioning



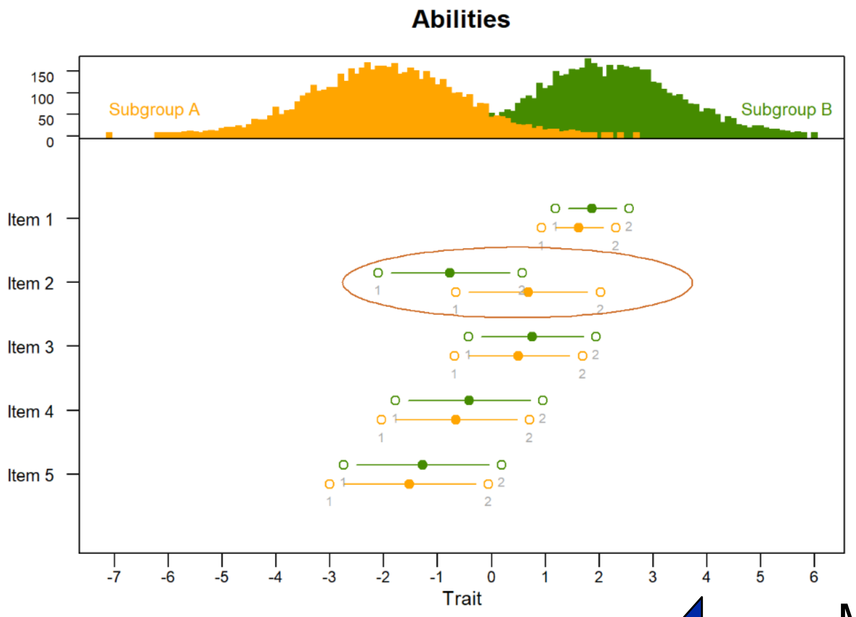
1) The ability of Subgroup A is lower than the ability of Subgroup B.

2) The difficulty of the item is similar for almost all items.

3) For a same level of ability, the difficulty of Item 2 differs across the Subgroup A and Subgroup B.

May not be a measurement bias or problem with the construct validity.

Differential Item Functioning?



1) The ability of Subgroup A is lower than the ability of Subgroup B.

2) The difficulty of the item is similar for almost all items.

May not be a measurement bias or problem with the construct validity.

3) For a same level of ability, the difficulty of Item 2 differs across the Subgroup A and Subgroup B.

DIF: Item 2 performs very differently in the two constructs.

DIF in Rasch Analysis

In Rasch analysis the residual matrix is tested for patterns that indicate systematic differences in responses across subgroups.

One approach is a two way analysis of variance (ANOVA) of the residuals.

Two way because of: (1) a DIF variable (age, gender, language, survey year…) and (2) score groups and their interaction.

The score groups represent a division of the total scores into equal-sized score groups.

Ideally the subgroup size should be between 30-50 persons.

A certain total score (example score = 2), is found only in one group.

Typically, the total score continuum would not be divided in to much more than 10 groups.

ID	Total Score	Score Group
1	1	
2	1	1
3	2	
4	3	
5	4	2
6	4	
7	5	
8	5	3
9	5	
10	6	
11	7	4
12	7	

Differential Item Functioning
Uniform vs Non-Uniform DIF

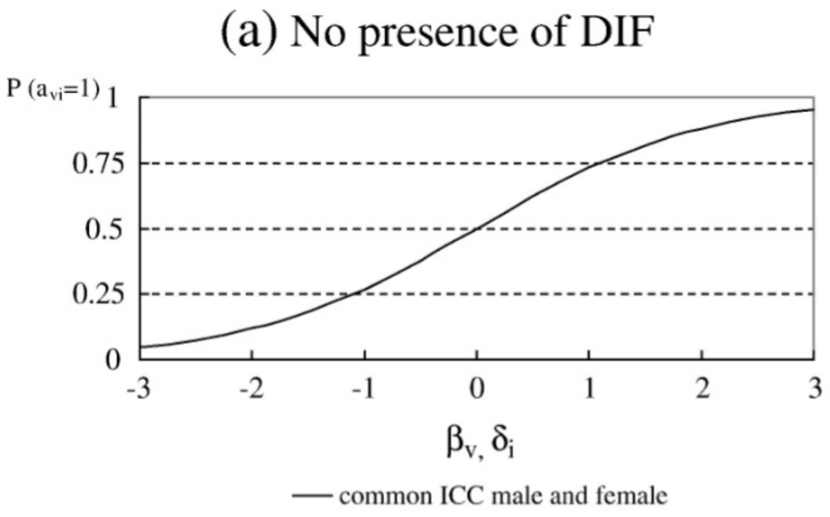
Uniform DIF:

Item difficulty estimates differ significantly across sample subgroups (age, gender, language, survey year, etc..)

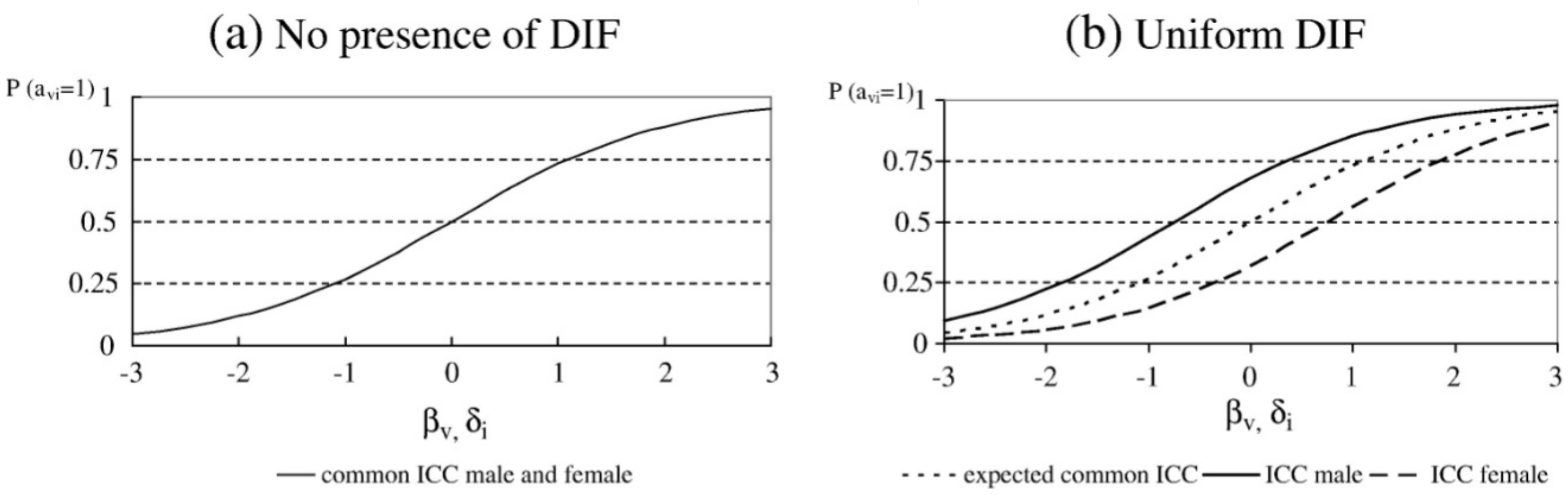
Non-Uniform DIF:

Item difficulty estimates differ significantly across sample subgroups and score level groups.

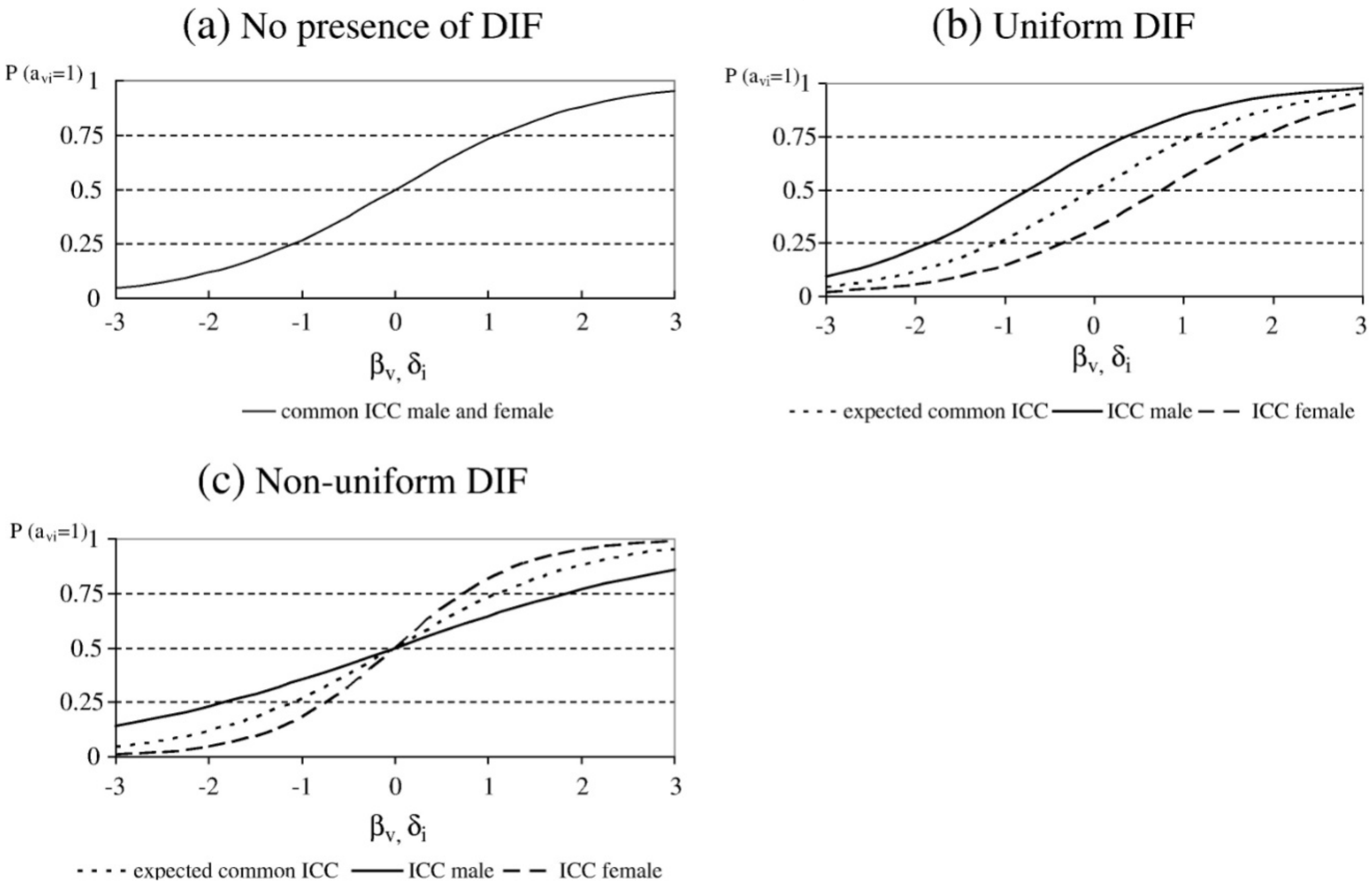
Differential Item Functioning



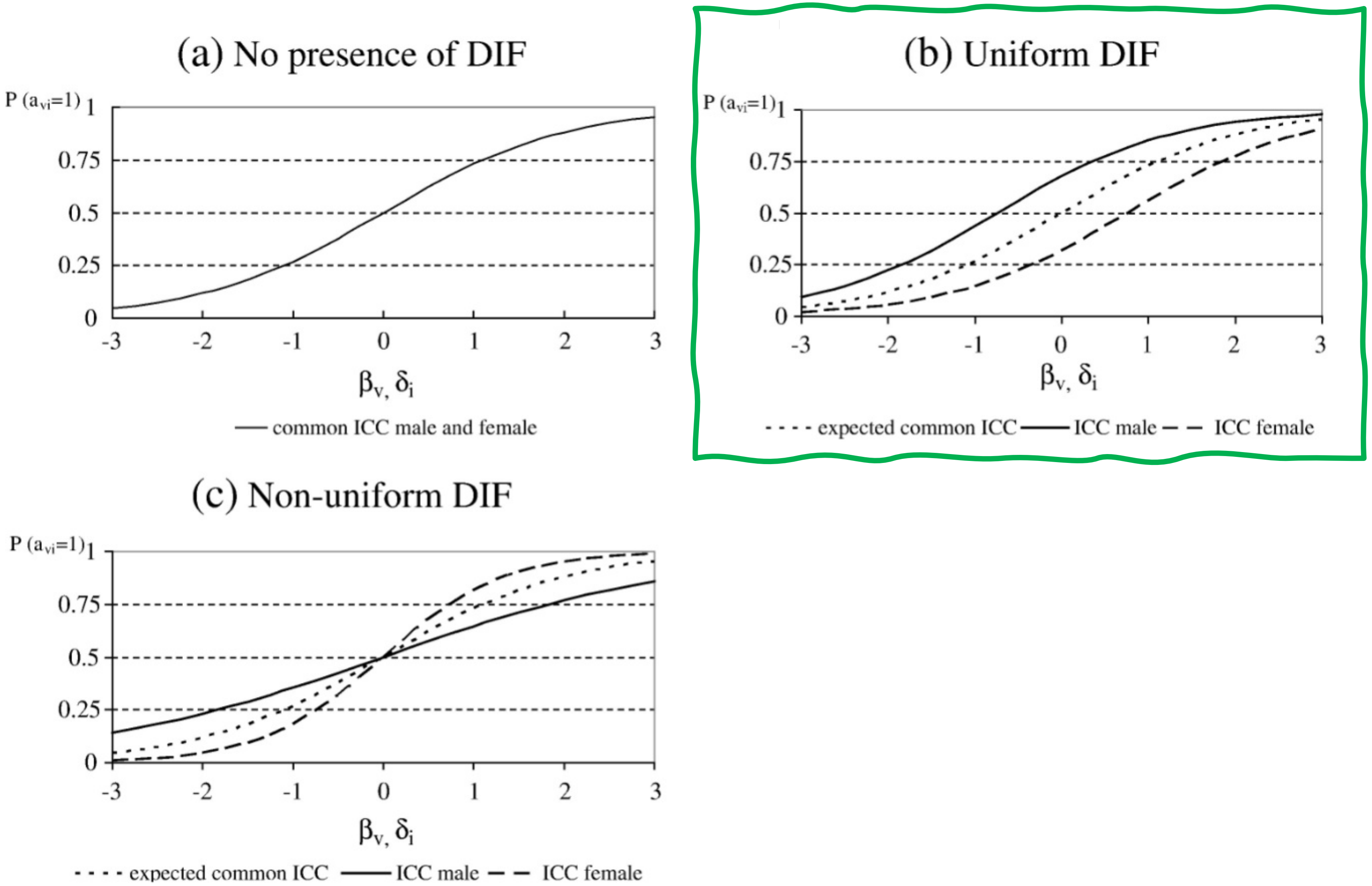
Differential Item Functioning



Differential Item Functioning

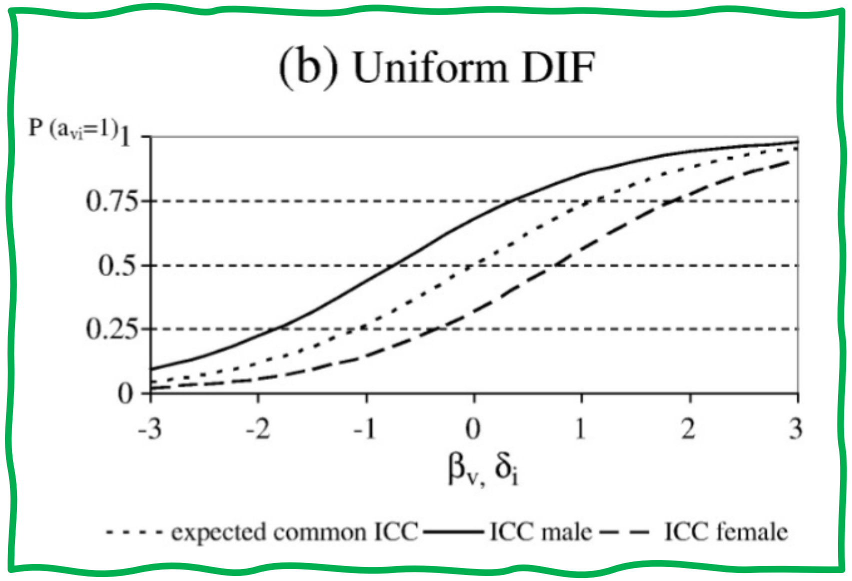


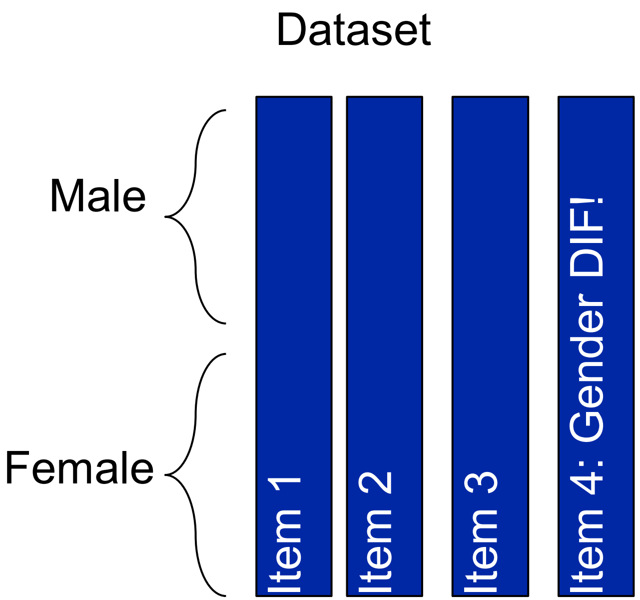
Differential Item Functioning



Differential Item Functioning Adjustment

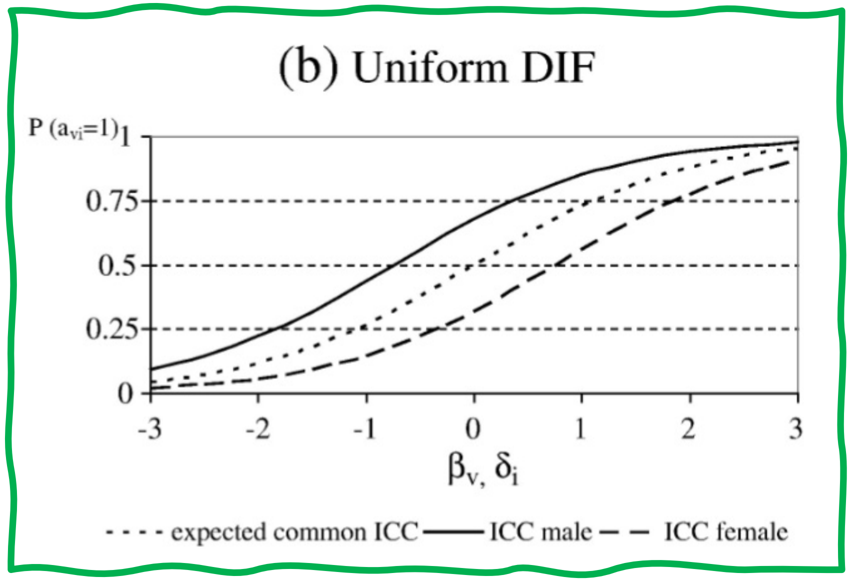
In presence of Uniform DIF, an approach to solve the DIF is item split.

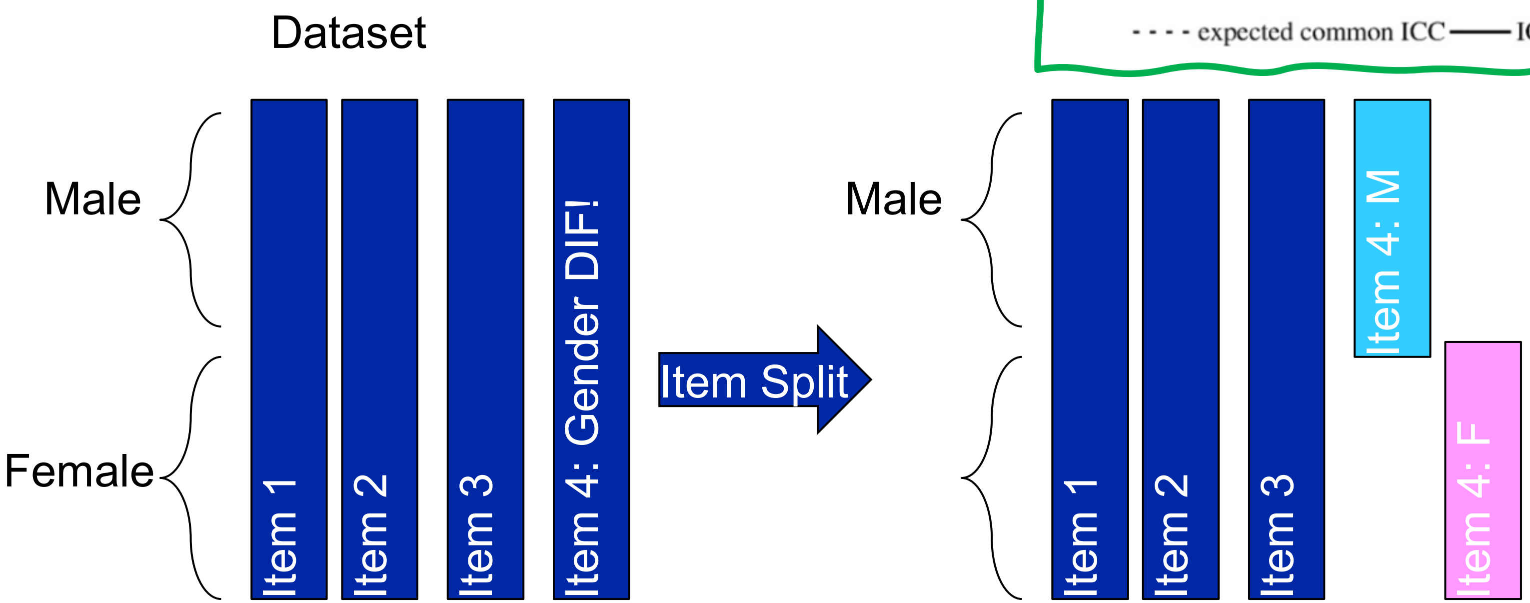




Differential Item Functioning Adjustment

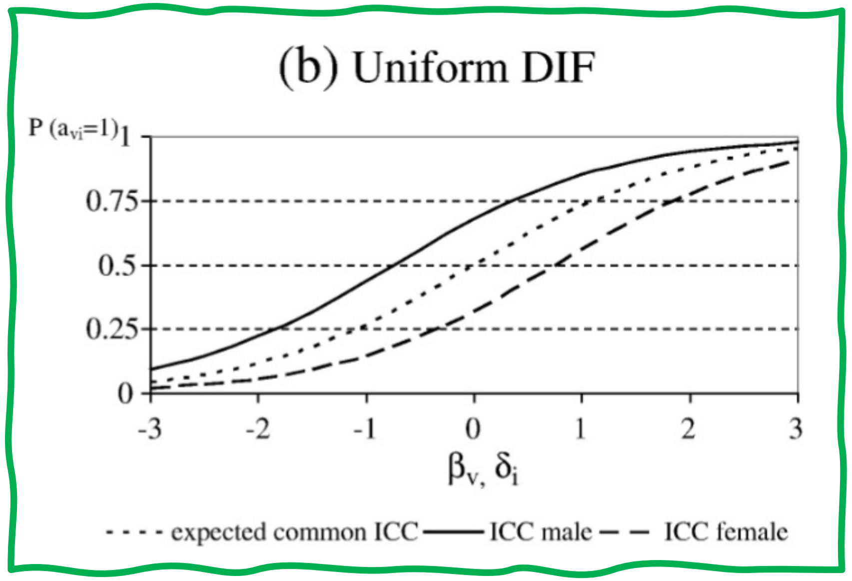
In presence of Uniform DIF, an approach to solve the DIF is item split.
Instead of one item, as many items as DIF subgroup level will enter the Rasch analysis.
This creates than adjusted subgroup specific metrics.

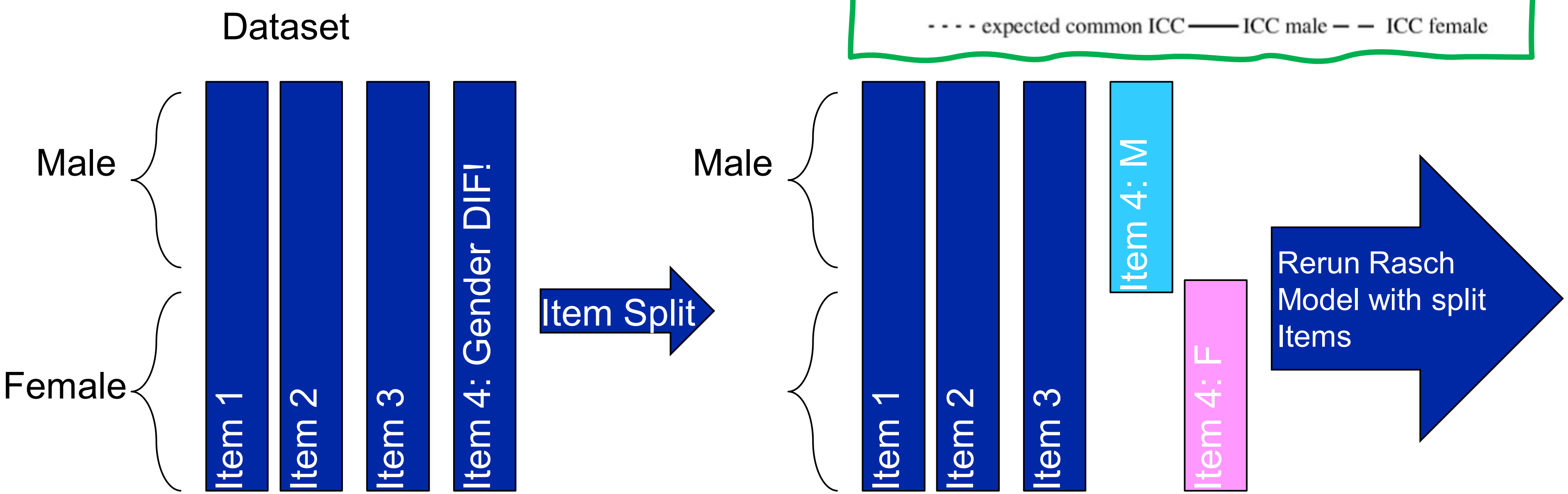




Differential Item Functioning Adjustment

In presence of Uniform DIF, an approach to solve the DIF is item split.
Instead of one item, as many items as DIF subgroup level will enter the Rasch analysis.
This creates than adjusted subgroup specific metrics.



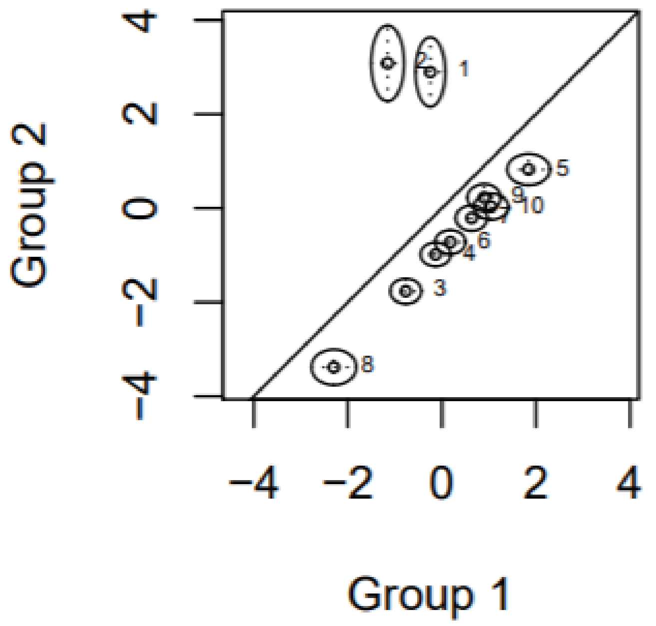


Differential Item Functioning Adjustment

Item **splitting is done stepwise**, starting with the item with the highest DIF.

Given the PCM estimation approach which centers the item difficulties to zero. It is not unusal, that the residuals may show artificial DIF in some principally DIF-free items as a reaction to items with very high DIF.

Example: items free of DIF should be on the diagonal. In the plot below no item is on the diagonal. Stepwise solving of DIF starting with the two upper items, is likely to recenter the items and show absence of DIF in a second run.



Group 1 = ex. male
Group 2 = ex. female

Rasch Analysis

A serie of assumptions have to be tested. If the scale ratings comply to these assumptions, the total score is interval-scaled.

- Stochastic ordering (fit of data to model)
- Monotonicity (ordering or response options)
- No local response dependencies or LID (no significant correlations between items)
- Unidimensionality (one latent construct)
- No differential item functioning or DIF (no sample subgroup effects)

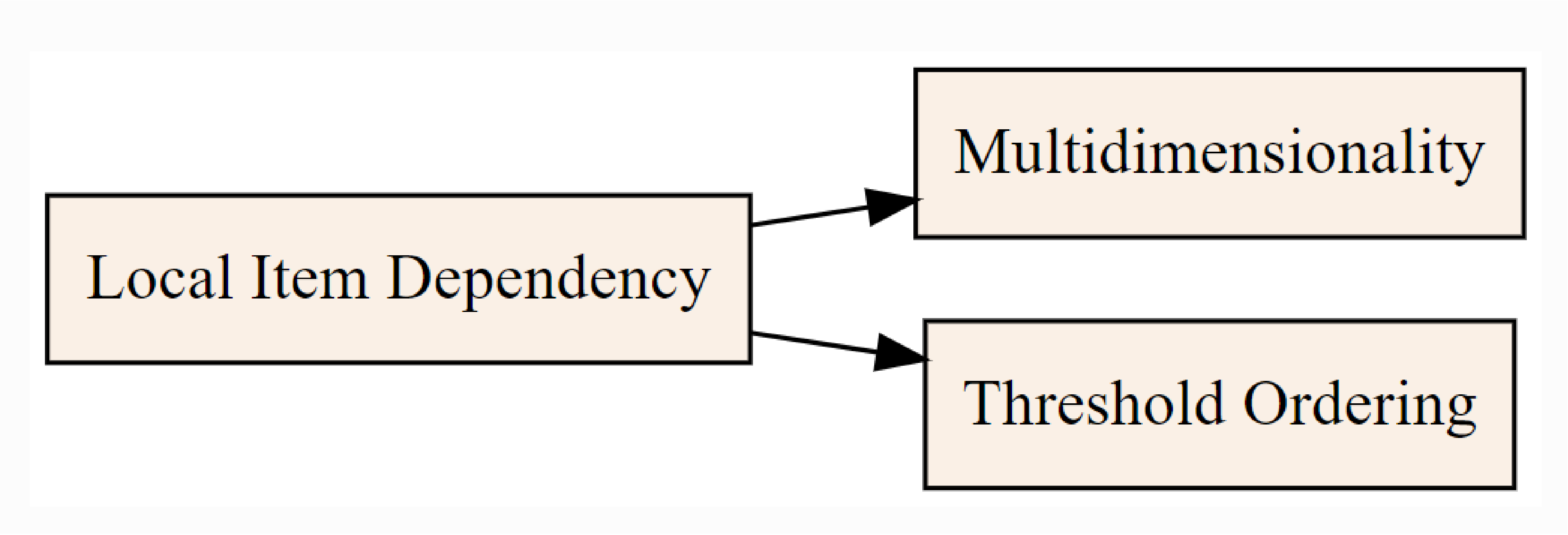
Rasch Analysis

Studies using Rasch analysis usually reports:

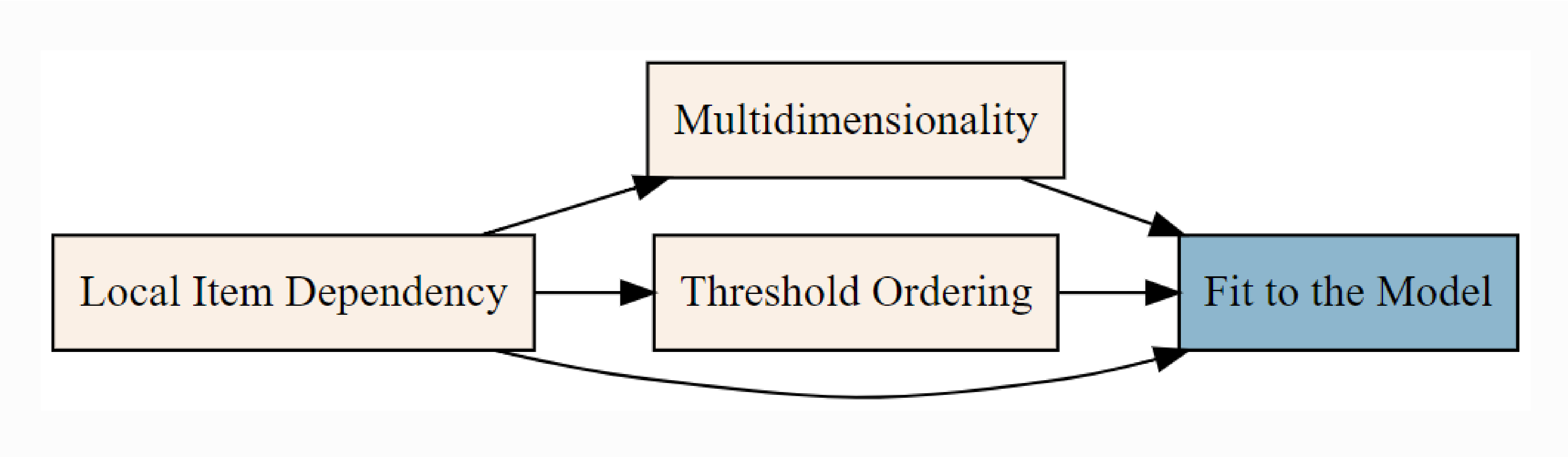
A) Fit statistics at start
B) Fit statistics when all breaches to the assumptions are fixed

Often the strategy to go from A to B is not reported.

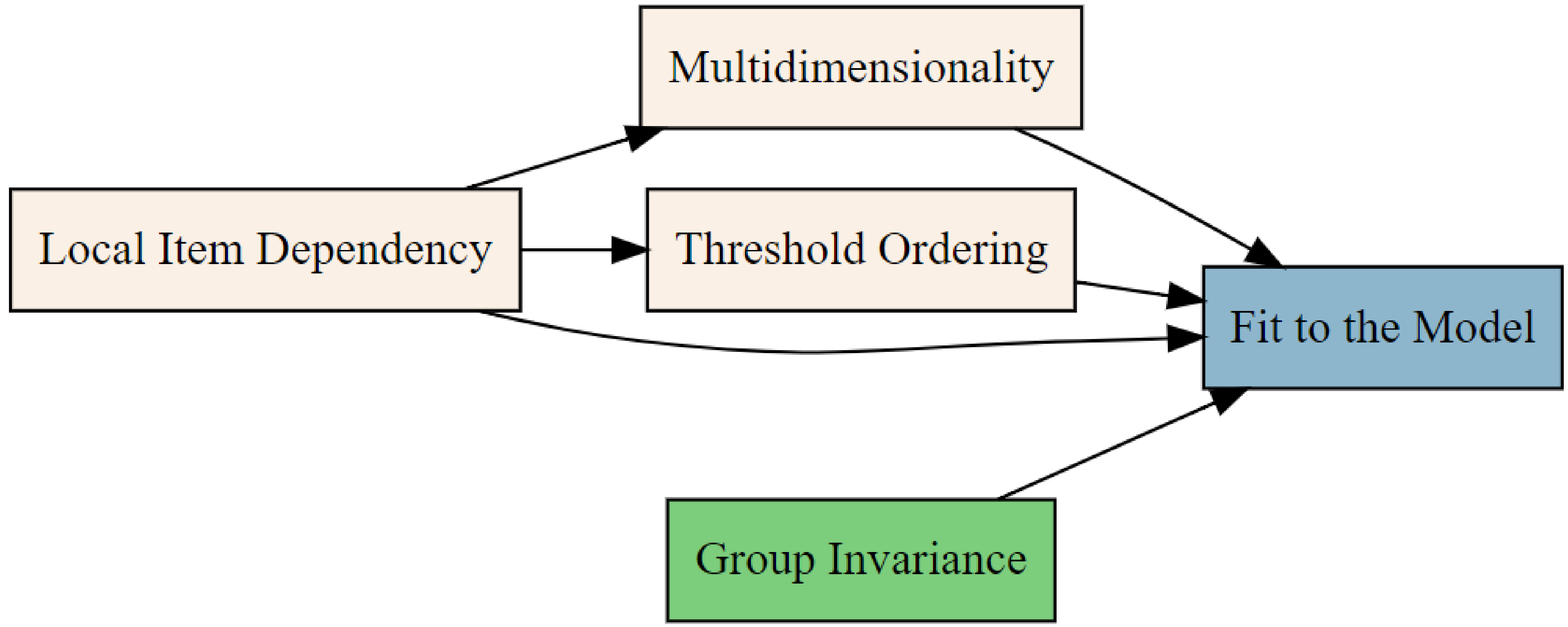
Rasch Analysis: Procedure



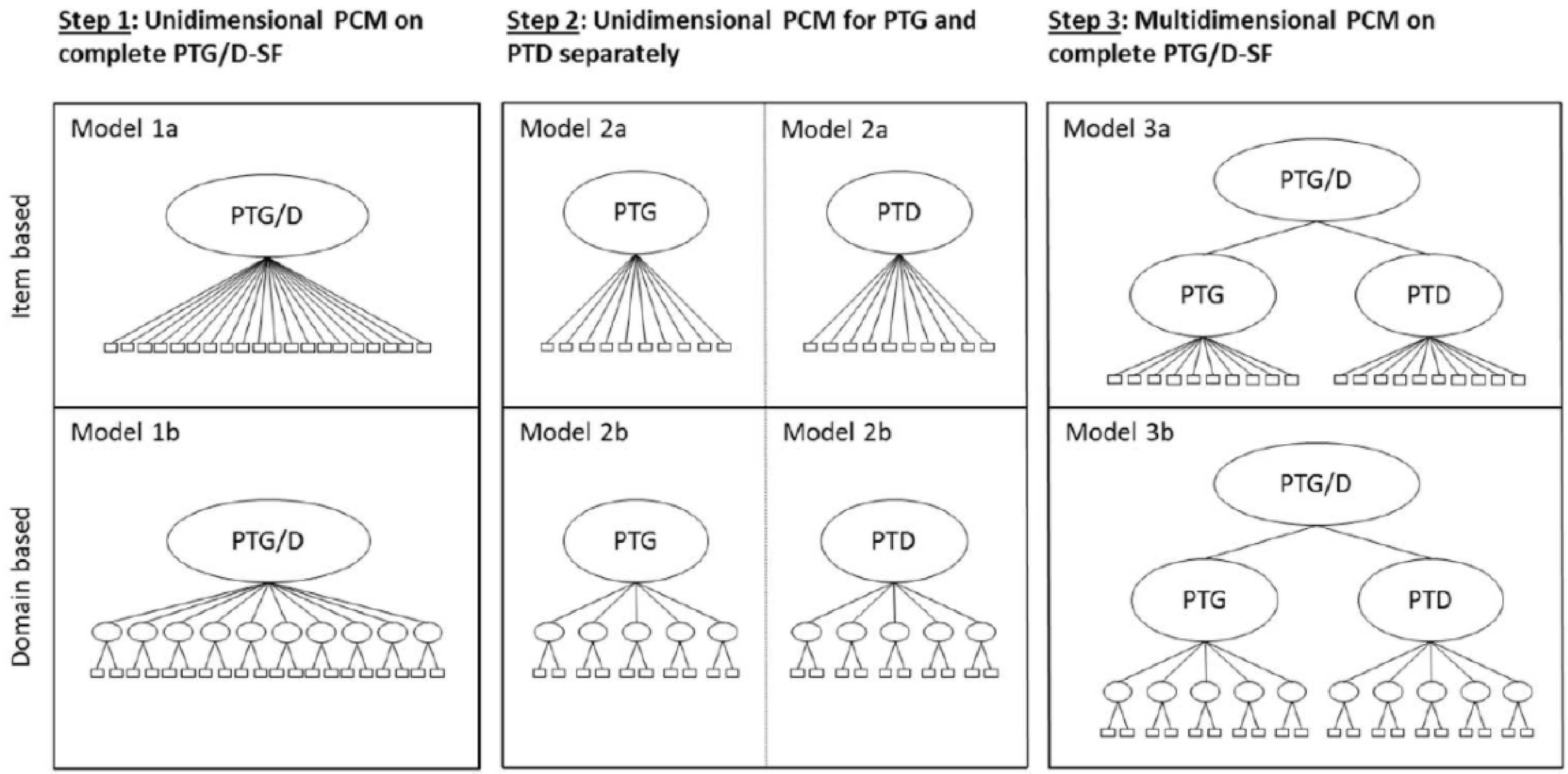
Rasch Analysis: Procedure



Rasch Analysis: Procedure



Rasch Analysis: Summarizing



Rasch Analysis: Summarizing

Table 4 Start and final model targeting fit of entire WHODAS 2.0, each subscale, and the calibration of domains as items

Dimension		Stage	Item difficulty		Person ability		Reliability		LID	Uniform DIF	Non-uniform DIF
			Mean	SD	Mean	SD	PSI	Cronbach alpha			
All	WHODAS 2.0	Start	0.05	0.71	−0.13	0.78	0.95	0.95	Yes	Yes	No
D1	Understanding and communicating	Start & Final	0.44	1.26	−0.58	1.34	0.91	0.91	No	No	No
D2	Getting around	Start	0.35	1.23	0.59	1.18	0.91	0.88	Yes	No	No
		Final	0.37	1.35	0.73	1.25	0.87	0.84	No	No	Yes
D3	Self-care	Start	0.54	1.90	−0.33	1.32	0.92	0.87	Yes	Yes	No
		Final	0.46	1.83	−0.36	1.11	0.89	0.67	No	Yes	No
D4	Getting along with people	Start	0.31	1.10	0.01	1.18	0.91	0.89	No	No	No
		Final	0.41	1.62	0.05	1.47	0.90	0.87	No	No	No
D5(1)	Household activities	Start & Final	2.15	5.00	2.39	4.04	0.98	0.99	No	No	No
D6	Participation in society	Start	0.25	0.73	0.26	1.01	0.90	0.88	Yes	Yes	No
		Final	0.26	0.93	0.27	1.05	0.89	0.83	No	Yes	No
Testlet		Start	0.02	0.96	−0.03	0.27	0.85	0.83	Yes	Yes	No
		Final	0.01	0.93	−0.02	0.22	0.79	0.75	No	Yes	No

PSI Person separation index, *LID* Local item dependency, *DIF* Differential item functioning