

Forced-choice Ranking

Sunthud Pornprasertmanit

Presentation Given at Applied Psychology Graduate Program,
Faculty of Psychology, Chulalongkorn University

Outline

- Methods of collecting forced-choice data
- Scaling single-stimulus items
- Apply single-stimulus technique to ranked items
- Thurstonian item response model
- How to design a test with ranking items

Types of Item Responses

- Single-stimulus response format
- Forced-choice response format

Methods for Collecting Forced-choice Data

Block Size	Binary Preference	Graded Preference	Proportion Preference
2	Paired Comparison	Graded Paired Comparison	Proportional Paired Comparison
≥ 3	Ranking Partial Ranking Q-sort	Blocked Graded Paired Comparison	Proportional-of-total amount task

Binary Preference

- Paired Comparison

Choose the statement that best describes you:

I am the life of the party
I prefer small gatherings

MBTI

Binary Preference

- Ranking

	H	L
I have highly original ideas.		X
I am prepared to tell people if they are mistaken.	X	
I am modest about what I have achieved		

SHL Personality Test

Binary Preference

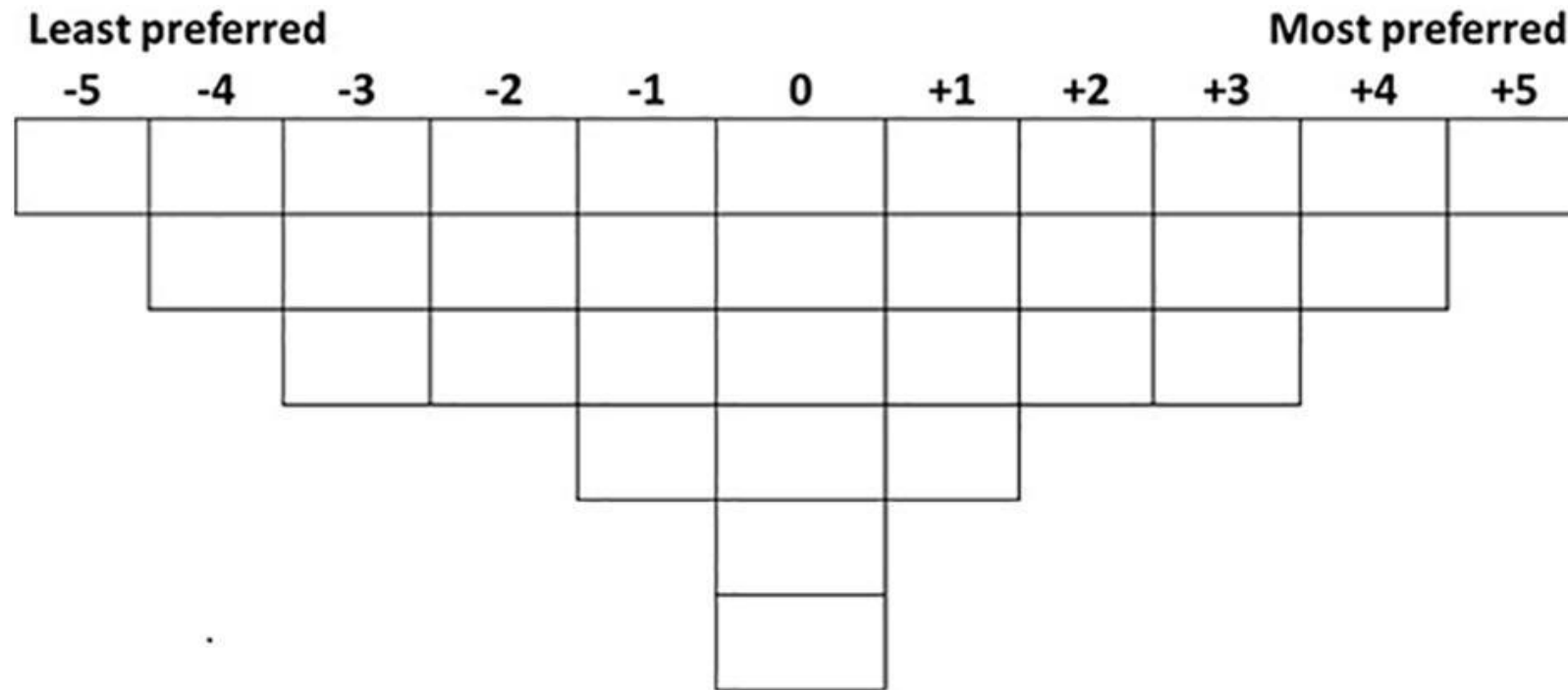
- Partial Ranking

	H	L
Has an excellent appetite		
Watches too much television		X
Follows a well-balanced diet	X	
Doesn't get enough exercise		

Gordon Personal Profile Inventory

Binary Preference

- Q-sort



Gao & Soranzo (2020)

Graded Preference

- Graded paired comparison

If I had to choose, I would say

	++	+	+	++	
Once I begin a project I finish it at all costs					I usually do what others recommend me to do

+ = a little more like me

++ = much more like me

Graded Preference

- Blocked Graded paired comparison

If I had to choose, I would say

	++	+	+	++	
A. Once I begin a project I finish it at all costs	X				B. I usually do what others recommend me to do
A. Once I begin a project I finish it at all costs		X			C. I often anticipate the needs of other
B. I usually do what others recommend me to do	X				C. I often anticipate the needs of other

+ = a little more like me

++ = much more like me

Adjusted from talenttoday

Proportional Preference

- Paired Comparison

	Points (100 in total)
I pay attention to details	33
I have a good word for everyone	67

Adjusted from Brown (2016)

Proportional Preference

- Proportional-of-total amount task

	Points (100 in total)
I pay attention to details	50
I change my mood a lot	35
I have a good word for everyone	10
I catch on to things quickly	5

Brown (2016)

Advantages of Forced-Choice Formats

- Reduce acquiescence bias, e.g., leniency, severity, central tendency, extremity
- Reduce lack of differentiation in ratings, e.g., halo effect such that rate with an overgeneralized perception of a characteristic
- For binary preference, reduce the different perceptions of scale labels
- Reduce faking, social desirability

Scaling single-stimulus items

- Classical test theory
- Factor analysis model
- Categorical Items
- Dominance/Ideal-point model

Scaling single-stimulus items

- Classical test theory

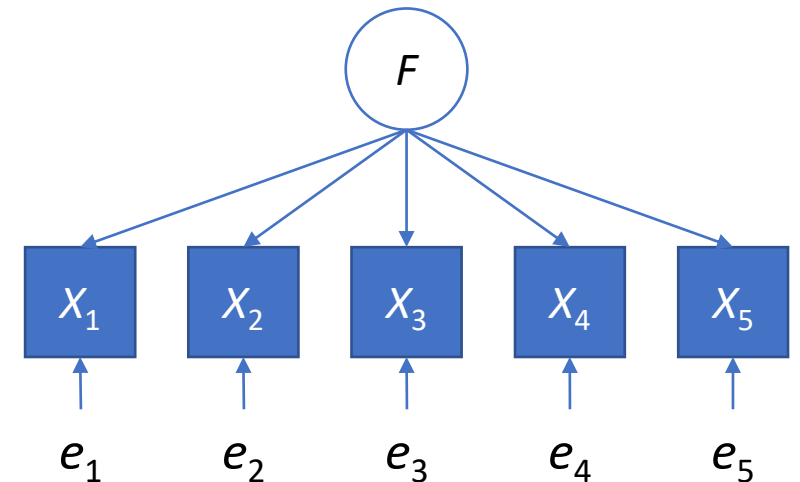
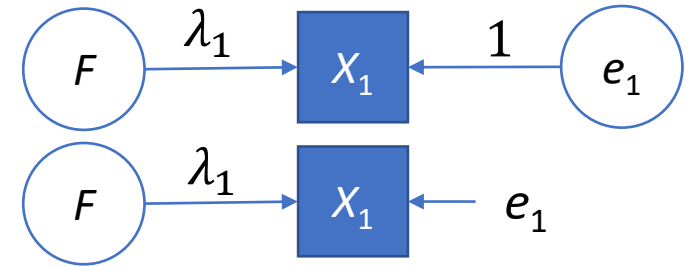
$$\begin{array}{rcl} X & = & T + e \\ X_1 & = & T + e_1 \\ X_2 & = & T + e_2 \\ X_3 & = & T + e_3 \\ X_4 & = & T + e_4 \\ & \vdots & \\ X_P & = & T + e_P \end{array} \qquad \bar{X} \approx T$$

Scaling single-stimulus items

- Factor Analysis Model

$$\begin{aligned} X_1 &= \lambda_1 F + e_1 \\ X_2 &= \lambda_2 F + e_2 \\ X_3 &= \lambda_3 F + e_3 \\ X_4 &= \lambda_4 F + e_4 \\ &\vdots \\ X_P &= \lambda_P F + e_P \end{aligned}$$
$$X = \lambda F + e$$

$$\bar{X} \approx F$$



$$X_1 = \lambda_{11}F_1 + e_1 = \lambda_{11}F_1 + 0F_2 + 0F_3 + e_1$$

$$X_2 = \lambda_{21}F_1 + e_2 = \lambda_{21}F_1 + 0F_2 + 0F_3 + e_2$$

$$X_3 = \lambda_{31}F_1 + e_3 = \lambda_{31}F_1 + 0F_2 + 0F_3 + e_3$$

$$X_4 = \lambda_{42}F_2 + e_4 = 0F_1 + \lambda_{42}F_2 + 0F_3 + e_4$$

$$X_5 = \lambda_{52}F_2 + e_5 = 0F_1 + \lambda_{52}F_2 + 0F_3 + e_5$$

$$X_6 = \lambda_{62}F_2 + e_6 = 0F_1 + \lambda_{62}F_2 + 0F_3 + e_6$$

$$X_7 = \lambda_{73}F_3 + e_7 = 0F_1 + 0F_2 + \lambda_{73}F_3 + e_7$$

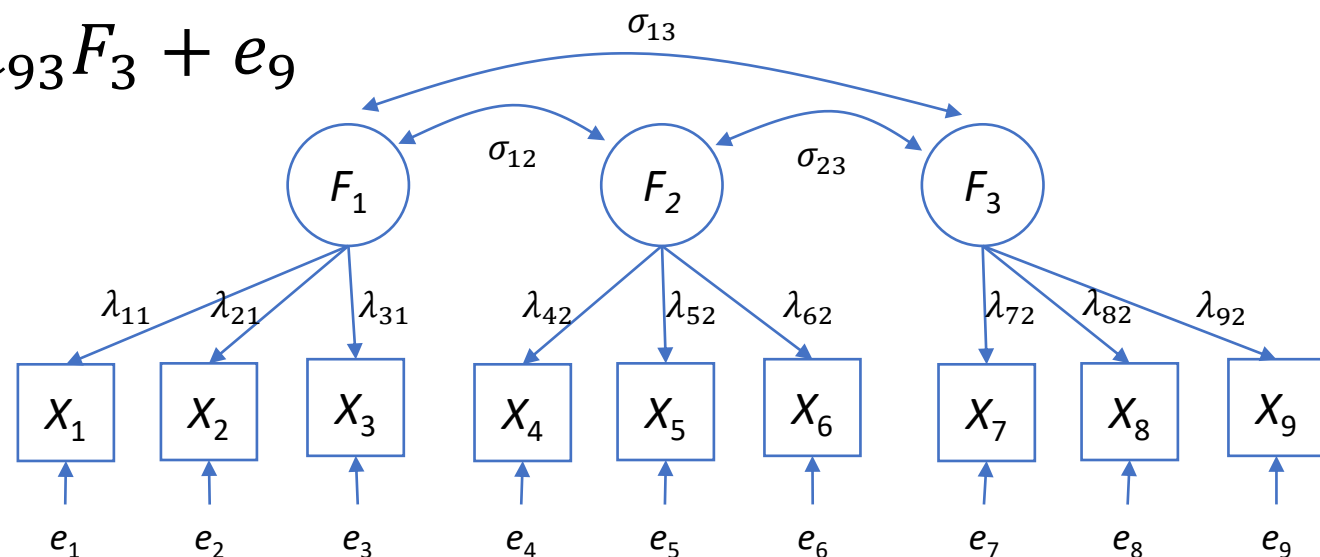
$$X_8 = \lambda_{83}F_3 + e_8 = 0F_1 + 0F_2 + \lambda_{83}F_3 + e_8$$

$$X_9 = \lambda_{93}F_3 + e_9 = 0F_1 + 0F_2 + \lambda_{93}F_3 + e_9$$

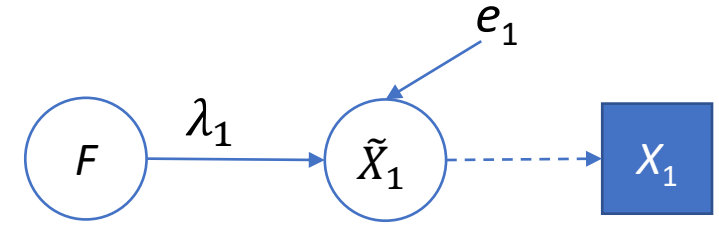
$$\frac{X_1 + X_2 + X_3}{3} \approx F_1$$

$$\frac{X_4 + X_5 + X_6}{3} \approx F_2$$

$$\frac{X_7 + X_8 + X_9}{3} \approx F_3$$



Scaling single-stimulus items

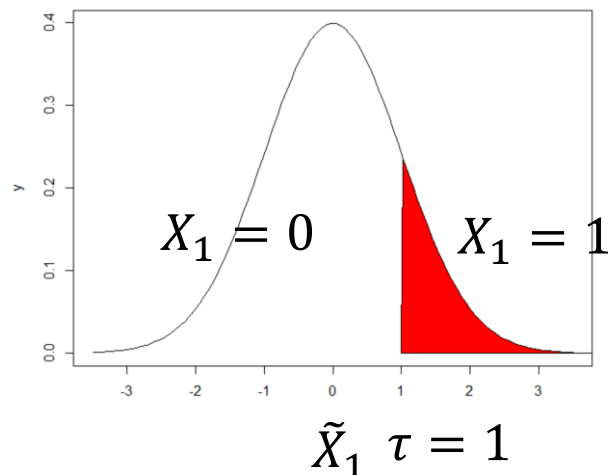


- Factor Analysis Model with Categorical Items

$$\tilde{X}_1 = \lambda_1 F + e_1$$

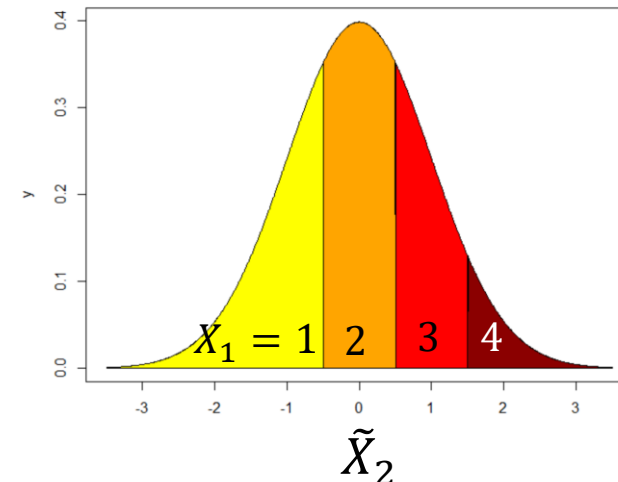
$$X_1 = \begin{cases} 1 & \text{if } \tau \leq \tilde{X}_1 < +\infty \\ 0 & \text{if } -\infty < \tilde{X}_1 < \tau \end{cases}$$

Probit Link



$$\tilde{X}_2 = \lambda_2 F + e_2$$

$$X_2 = \begin{cases} 1 & \text{if } -\infty < \tilde{X}_2 < \tau_1 \\ 2 & \text{if } \tau_1 \leq \tilde{X}_2 < \tau_2 \\ 3 & \text{if } \tau_2 \leq \tilde{X}_2 < \tau_3 \\ 4 & \text{if } \tau_3 \leq \tilde{X}_2 < +\infty \end{cases}$$



$$\begin{aligned} \tau_1 &= -0.5 \\ \tau_2 &= 0.5 \\ \tau_3 &= 1.5 \end{aligned}$$

$$\tilde{X}_1 = \lambda_{11}F_1 + e_1$$

$$\tilde{X}_2 = \lambda_{21}F_1 + e_2$$

$$\tilde{X}_3 = \lambda_{31}F_1 + e_3$$

$$\tilde{X}_4 = \lambda_{42}F_2 + e_4$$

$$\tilde{X}_5 = \lambda_{52}F_2 + e_5$$

$$\tilde{X}_6 = \lambda_{62}F_2 + e_6$$

$$\tilde{X}_7 = \lambda_{73}F_3 + e_7$$

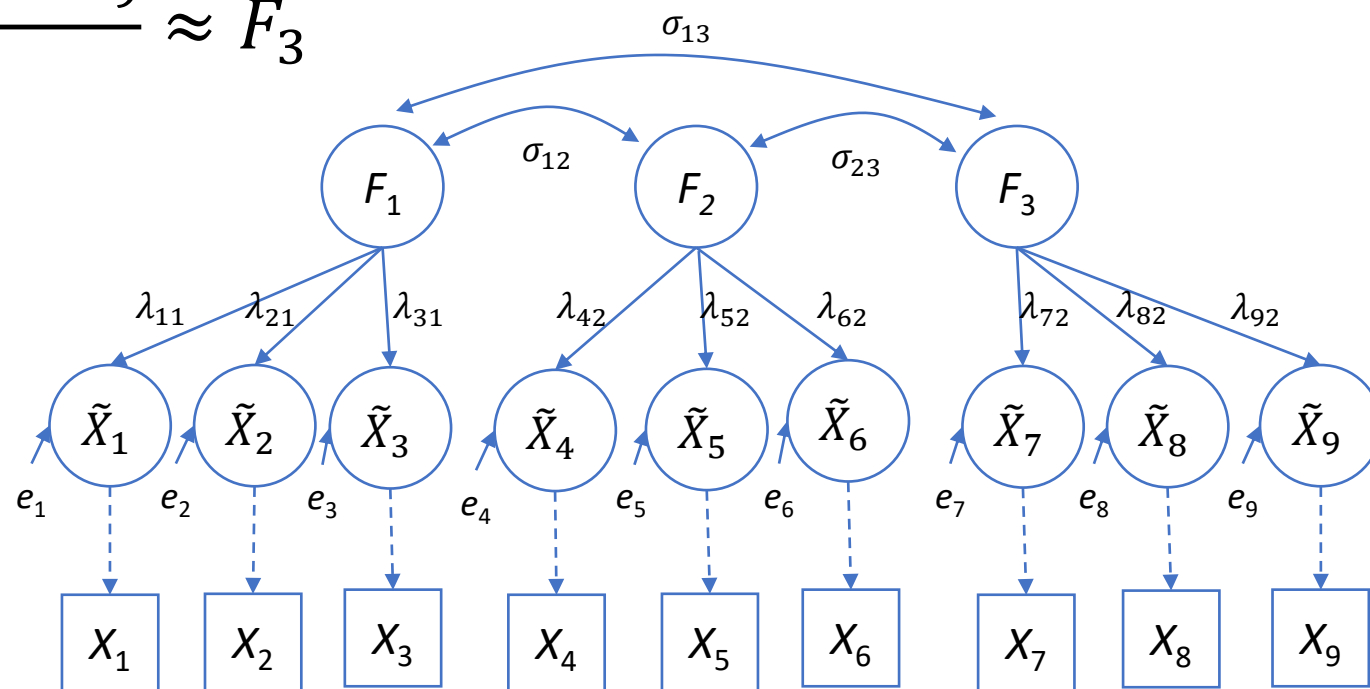
$$\tilde{X}_8 = \lambda_{83}F_3 + e_8$$

$$\tilde{X}_9 = \lambda_{93}F_3 + e_9$$

$$\frac{X_1 + X_2 + X_3}{3} \approx F_1$$

$$\frac{X_4 + X_5 + X_6}{3} \approx F_2$$

$$\frac{X_7 + X_8 + X_9}{3} \approx F_3$$



Scaling single-stimulus items

- Dominance vs. Ideal-point model

	Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree
I am the life of the party					
I like to be alone					
I prefer small gatherings					
I go to parties occasionally					

Scaling single-stimulus items

- Dominance vs. Ideal-point model

	Extreme Introvert	Introvert	Neutral	Extrovert	Extreme Extrovert
I am the life of the party	Strongly Disagree	Strongly Disagree	Disagree	Agree	Strongly Agree
I like to be alone	Strongly Agree	Agree	Disagree	Strongly Disagree	Strongly Disagree
I prefer small gatherings	Agree	Strongly Agree	Agree	Disagree	Strongly Disagree
I go to parties occasionally	Disagree	Agree	Strongly Agree	Agree	Disagree

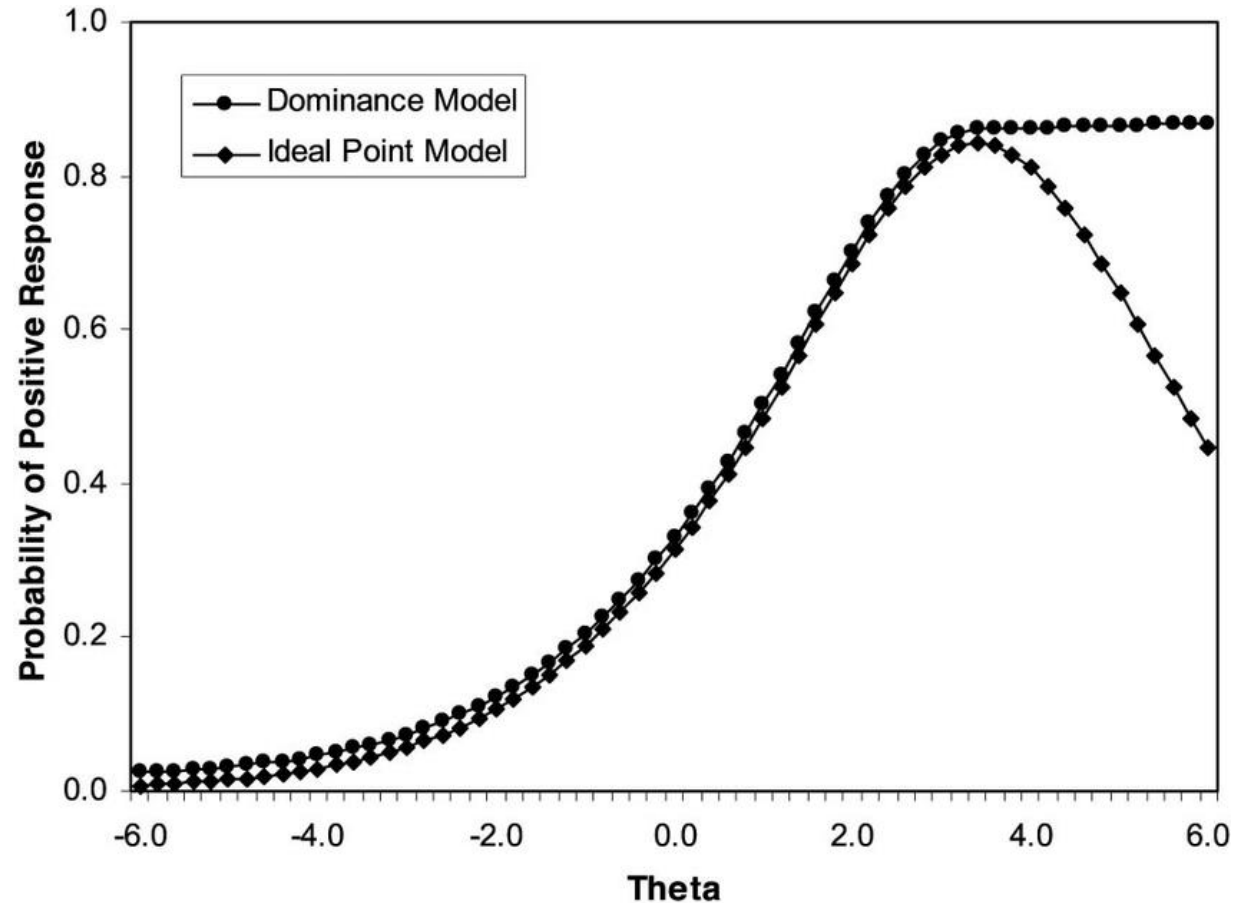
Scaling single-stimulus items

- Dominance Model

$$\tilde{X}_p = \mu_p + \lambda_p F + e_p$$

- Ideal-point model

$$\tilde{X}_p = \mu_p - \lambda_p |F - location_p| + e_p$$



Apply single-stimulus technique to ranked items

- Let F_1, F_2, F_3 , and F_4 be independent factors
- Simulate data of four factors and rank the factor scores within each person

F1	F2	F3	F4
0.8	-0.4	1.0	1.1
1.0	0.1	0.9	0.4
0.5	0.7	-0.5	-0.5
1.3	-0.9	-1.0	0.3
-1.3	-0.1	1.5	-1.2

R1	R2	R3	R4
3	4	2	1
3	1	2	4
2	1	4	3
2	3	1	4
4	1	3	2

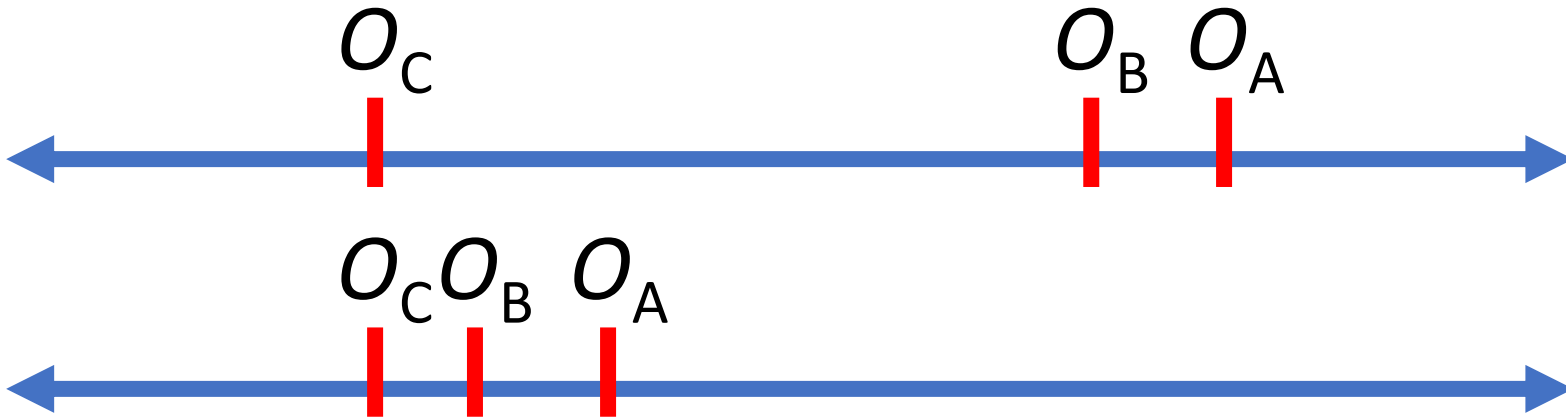
Types of Inference

- Interpersonal – scores are compared between individuals.
- Intrapersonal – scores are compared within individuals.
- Rank items are good for intrapersonal comparison but not so good for interpersonal comparison.

I am the sort of person who...

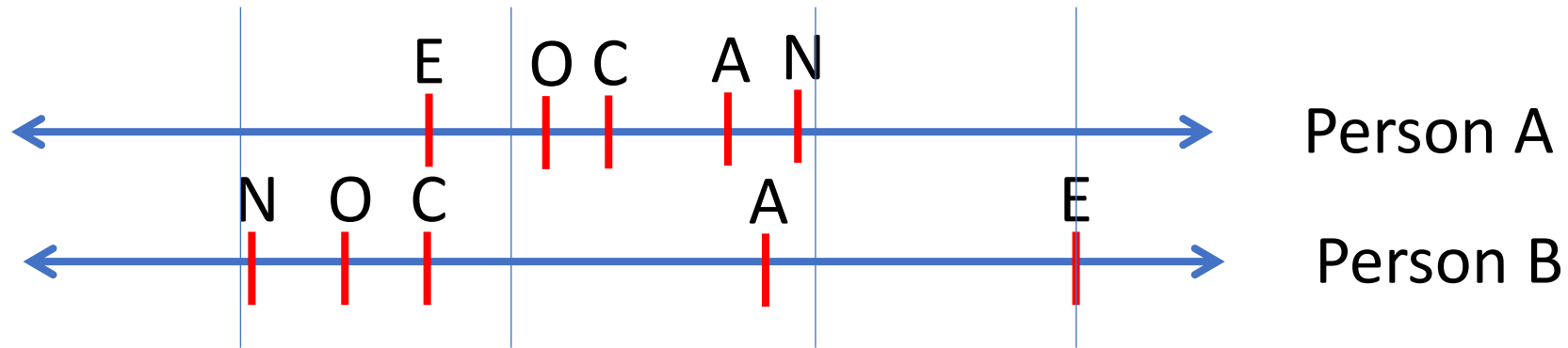
			M	L
1	A	I try out new activities.	☒	☐
	B	I consider other people's feelings.	☐	☐
	C	I like to understand the underlying theory.	☐	☒

From above, the degree of agreement in each statement (O_A , O_B , O_C) can be anywhere as long as the order is correct



Types of Inference

- For example,



Use single-stimulus items

	O	C	E	A	N
A	2.2	2.4	1.8	2.7	2.9
B	1.4	1.7	4.0	2.7	1.1

Use ranked items

	O	C	E	A	N
A	2	3	1	4	5
B	2	3	5	4	1

Using ranks to compare individuals is misleading.

Apply single-stimulus technique to ranked items

- We found that the correlations between ranks are negative (-.33 for correlations between ranks of 4 independent factors)
- Even if the correlations between factors are positive, the correlations between ranks are also negative.
- The negative correlations between ranks is called ipsative.
- Thus, classical test theory, factor analysis, or other methods for single-stimulus items are not appropriate.
- As the number of compared statements increases, the negative correlations between ranks are closer to 0.
 - Thus, old literature recommend comparing many traits (i.e., 30 or more traits).

Apply single-stimulus technique to ranked items

- In general, ranking is used for alleviating faking effects, so the social desirability of each statement is equal. This is called equally keyed items.
- Unequally keyed items are the ranking that desirable and undesirable items are compared. This type of rank item is fakable.
- Unequally keyed items, especially from a single construct, have similar properties to single-stimulus items.
 - Thus, including them will alleviate ipsativity.
 - However, the ipsative problem still exists so we call partial ipsative.

Thurstonian Item Response Theory Model

- Thurstonian IRT model introduced by Brown and Maydeu-Olivares (2011, 2013, 2016)
- They proposed how multiple factors affects how persons decide in the binary preference (and other types of comparison).
 - Originally, Thurstone's (1927) law of comparative judgment showed that the decision of binary preference data between statement 1 and 2:

$$\text{Selection} \left\{ \begin{array}{l} 1 \text{ if utility of 1} > \text{utility of 2} \\ 2 \text{ if utility of 2} > \text{utility of 1} \end{array} \right.$$

Let t_1 and t_2 be the utility of Statement 1 and 2:

$$i_{\{1,2\}} = \begin{cases} 1 & \text{if } t_1 \geq t_2 \\ 0 & \text{if } t_1 < t_2 \end{cases}$$

t_1 and t_2 are similar to latent continuous variables underlying categorical items in categorical factor analysis model.

Transform ranking to binary preference

1

		M	L
1	I try out new activities.	☒	☐
2	I consider other people's feelings.	☐	☐
3	I like to understand the underlying theory.	☐	☒

$$i_{1,2} = 1, i_{1,3} = 1, i_{2,3} = 1$$

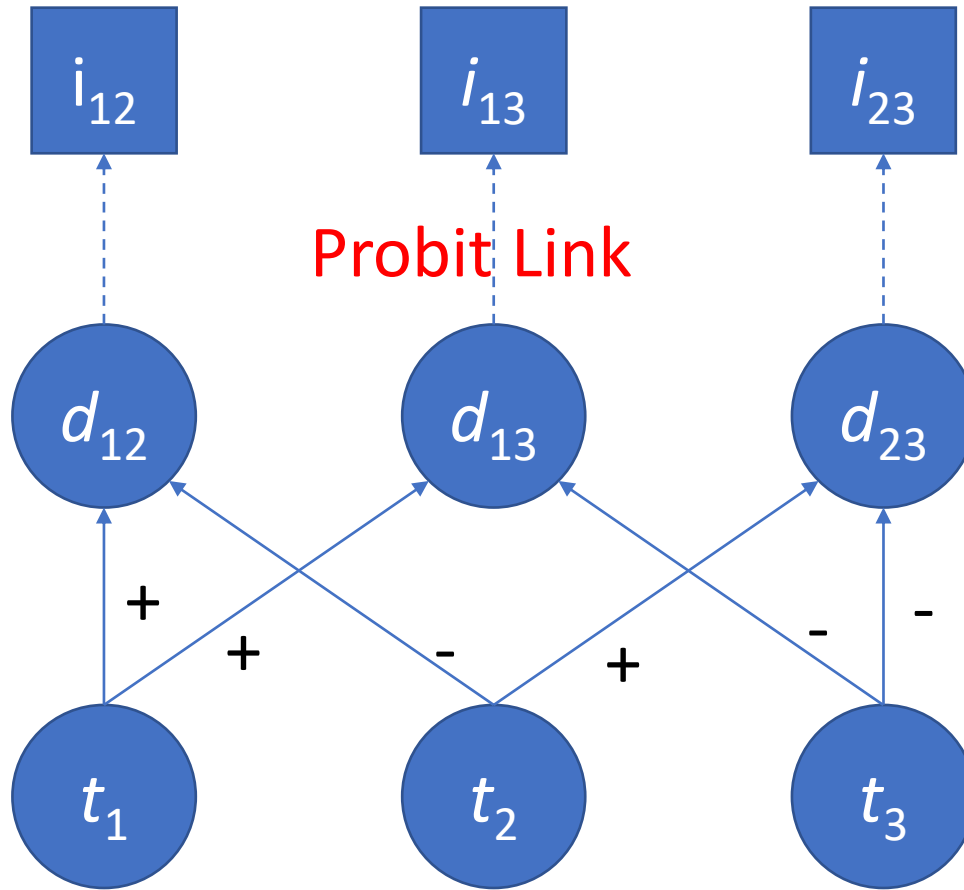
	Rank (1=Highest)
1. I have highly original ideas.	3
2. I am prepared to tell people if they are mistaken.	1
3. I am modest about what I have achieved	4
4. I make friends easily	2

$$i_{1,2} = 0, i_{1,3} = 1, i_{1,4} = 0, i_{2,3} = 1, i_{2,4} = 1, i_{3,4} = 0$$

Transform ranking to binary preference

	H	L
1. Has an excellent appetite		
2. Watches too much television		X
3. Follows a well-balanced diet	X	
4. Doesn't get enough exercise		

$$i_{1,2} = 1, i_{1,3} = 0, i_{1,4} = \text{NA}, i_{2,3} = 0, i_{2,4} = 0, i_{3,4} = 1$$



$$d_{12} = t_1 - t_2$$

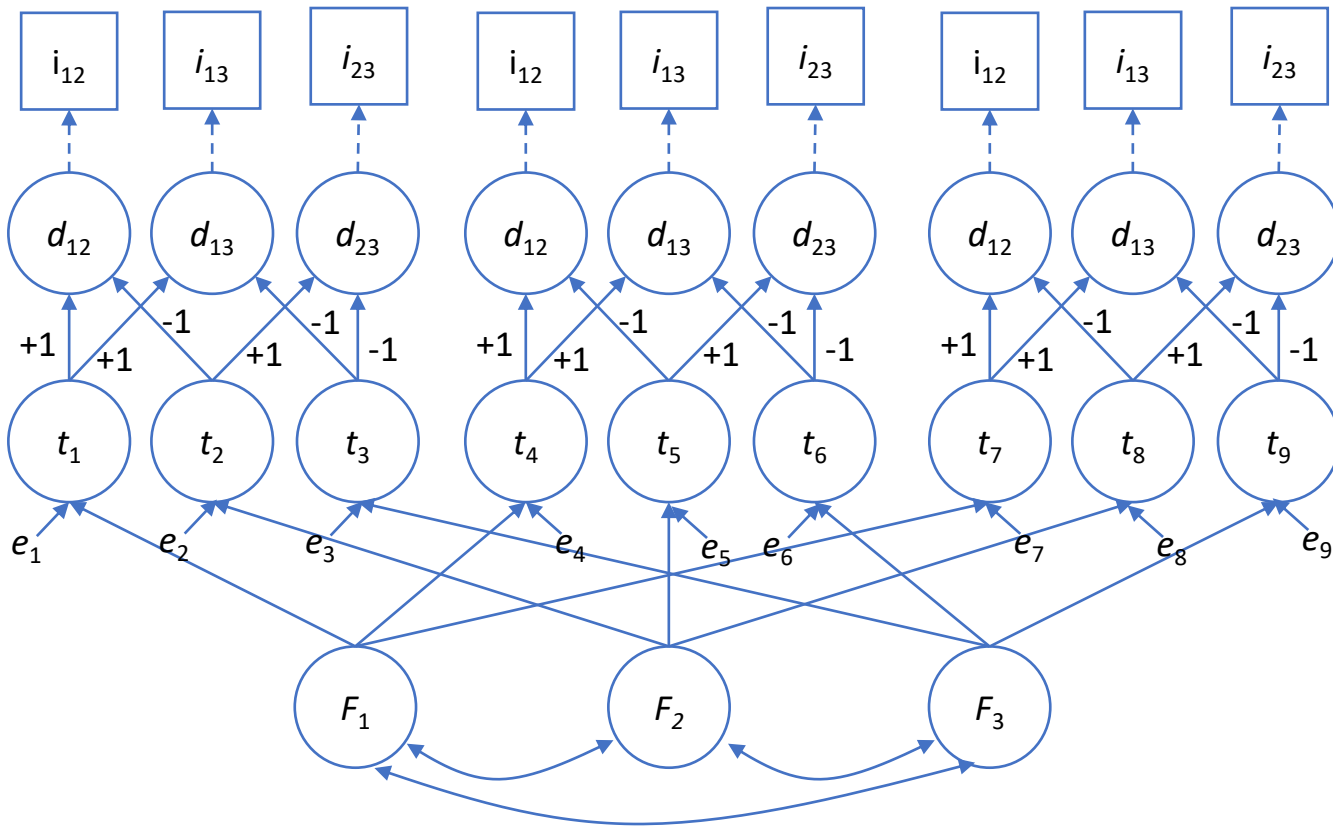
$$d_{13} = t_1 - t_3$$

$$d_{23} = t_2 - t_3$$

$$i_{12} = \begin{cases} 1 & \text{if } d_{12} \geq 0 \\ 0 & \text{if } d_{12} < 0 \end{cases}$$

$$i_{13} = \begin{cases} 1 & \text{if } d_{13} \geq 0 \\ 0 & \text{if } d_{13} < 0 \end{cases}$$

$$i_{23} = \begin{cases} 1 & \text{if } d_{23} \geq 0 \\ 0 & \text{if } d_{23} < 0 \end{cases}$$



$$\begin{aligned}
 t_1 &= \lambda_1 F_1 + e_1 & d_{12} &= t_1 - t_2 \\
 t_2 &= \lambda_2 F_2 + e_2 & d_{13} &= t_1 - t_3 \\
 t_3 &= \lambda_3 F_3 + e_3 & d_{23} &= t_2 - t_3 \\
 t_4 &= \lambda_4 F_1 + e_4 & d_{45} &= t_4 - t_5 \\
 t_5 &= \lambda_5 F_2 + e_5 & d_{46} &= t_4 - t_6 \\
 t_6 &= \lambda_6 F_3 + e_6 & d_{56} &= t_5 - t_6 \\
 t_7 &= \lambda_7 F_1 + e_7 & d_{78} &= t_7 - t_8 \\
 t_8 &= \lambda_8 F_2 + e_8 & d_{79} &= t_7 - t_9 \\
 t_9 &= \lambda_9 F_3 + e_9 & d_{89} &= t_8 - t_9
 \end{aligned}$$

$$i_{jk} = \begin{cases} 1 & \text{if } d_{jk} \geq 0 \\ 0 & \text{if } d_{jk} < 0 \end{cases}$$

This is Thurstonian IRT Model.

This is called second-order factor model form.

Thurstonian IRT Model

- The second-order model cannot get factor score estimates because of the complexity of the model.
- Brown and Maydeu-Olivares (2013) showed that the second-order factor model can collapse to one-level factor analysis model.
 - First, substitute t in the d definition equation
 - Second, calculate the residual variance of each d , as well as the covariance of d s from the same rank item

$$d_{12} = (\lambda_1 F_1 + e_1) - (\lambda_2 F_2 + e_2)$$

$$d_{13} = (\lambda_1 F_1 + e_1) - (\lambda_3 F_3 + e_3)$$

$$d_{23} = (\lambda_2 F_2 + e_2) - (\lambda_3 F_3 + e_3)$$

$$d_{45} = (\lambda_4 F_1 + e_4) - (\lambda_5 F_2 + e_5)$$

$$d_{46} = (\lambda_4 F_1 + e_4) - (\lambda_6 F_3 + e_6)$$

$$d_{56} = (\lambda_5 F_2 + e_5) - (\lambda_6 F_3 + e_6)$$

$$d_{78} = (\lambda_7 F_1 + e_7) - (\lambda_8 F_2 + e_8)$$

$$d_{79} = (\lambda_7 F_1 + e_7) - (\lambda_9 F_3 + e_9)$$

$$d_{89} = (\lambda_8 F_2 + e_8) - (\lambda_9 F_3 + e_9)$$

$$d_A = \lambda_{A1} F_1 - \lambda_{A2} F_2 + e_A$$

$$d_B = \lambda_{B1} F_1 - \lambda_{B3} F_3 + e_B$$

$$d_C = \lambda_{C1} F_2 - \lambda_{C3} F_3 + e_C$$

$$d_D = \lambda_{D1} F_1 - \lambda_{D2} F_2 + e_D$$

$$d_E = \lambda_{E1} F_1 - \lambda_{E3} F_3 + e_E$$

$$d_F = \lambda_{F1} F_2 - \lambda_{F3} F_3 + e_F$$

$$d_G = \lambda_{G1} F_1 - \lambda_{G2} F_2 + e_G$$

$$d_H = \lambda_{H1} F_1 - \lambda_{H3} F_3 + e_H$$

$$d_I = \lambda_{I1} F_2 - \lambda_{I3} F_3 + e_I$$

$$Var(e_A) = Var(e_1 - e_2) = Var(e_1) + Var(e_2)$$

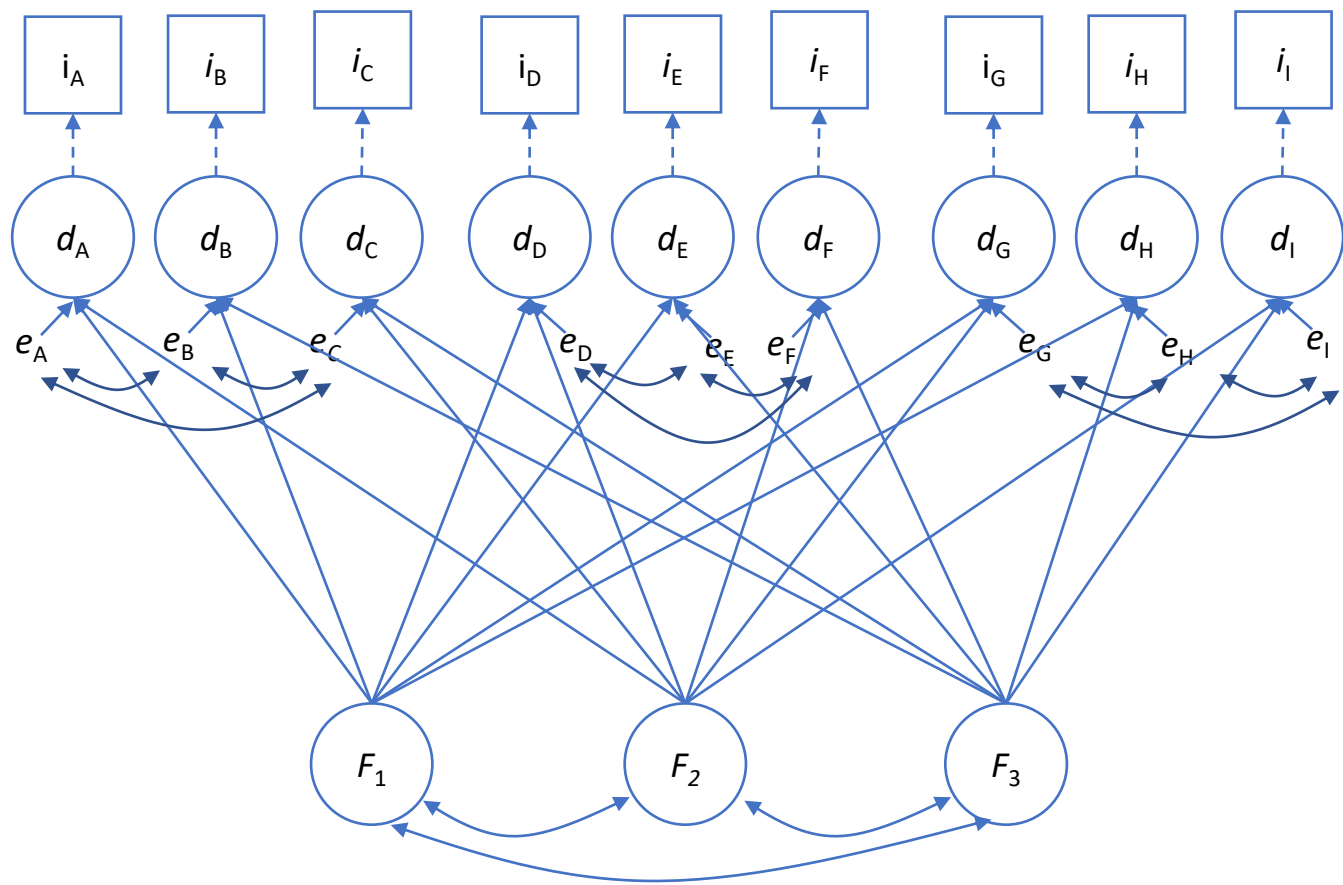
$$Var(e_B) = Var(e_1 - e_3) = Var(e_1) + Var(e_3)$$

$$Var(e_C) = Var(e_2 - e_3) = Var(e_2) + Var(e_3)$$

$$Cov(e_A, e_B) = Cov(e_1 - e_2, e_1 - e_3) = Cov(e_1, e_1) + Cov(e_1, -e_3) + Cov(-e_2, e_1) + Cov(-e_2, -e_3) = Var(e_1)$$

$$Cov(e_A, e_C) = Cov(e_1 - e_2, e_2 - e_3) = Cov(e_1, e_2) + Cov(e_1, -e_3) + Cov(-e_2, e_2) + Cov(-e_2, -e_3) = -Var(e_2)$$

$$Cov(e_B, e_C) = Cov(e_1 - e_3, e_2 - e_3) = Cov(e_1, e_2) + Cov(e_1, -e_3) + Cov(-e_3, e_2) + Cov(-e_3, -e_3) = Var(e_3)$$



Thurstonian IRT Model

- It is complicated to write codes manually, especially in a large number of rank item.
- Brown provided Excel macro to automatically generate Mplus code
- It is necessary to ensure that you set the binary preference dataset correctly. It is so tedious and error-prone to transform from ranks to binary preference.
- Bürkner, also, wrote R package to run TIRT model using Mplus, lavaan (frequentist analysis), or Stan (Bayesian analysis)

Example Dataset 1

- Three factors
 - Factor 1: Item 1, 4, 7, 10, 13, 16, 19, 22, 25, 28
 - Factor 2: Item 2, 5, 8, 11, 14, 17, 20, 23, 26, 29
 - Factor 3: Item 3, 6, 9, 12, 15, 18, 21, 24, 27, 30
- Positive Items: 1-15, Negative Items: 16-30
- V1-V30: continuous items
- C1-C30: categorical items with 4 categories
- rank11, rank12, rank13 are three statements in the rank item 1 (using item 1, 2, 3) (1 = most like me, 3 = least like me)
- g1_2 is graded preference comparison between item1 and item2 (1,2 = prefer item 1, 3,4 = prefer item 2)

TIRT : Graded Preference

- Binary preference categorizes d to two groups:

$$i_{12} = \begin{cases} 1 & \text{if } d_{12} \geq 0 \text{ or } t_1 \geq t_2 \\ 0 & \text{if } d_{12} < 0 \text{ or } t_1 < t_2 \end{cases}$$

- Graded preference simply categorizes d into multiple groups:

$$i_{12} = \begin{cases} 1 & \text{if } -\infty < d_{12} \leq \tau_1 \\ 2 & \text{if } \tau_1 < d_{12} \leq \tau_2 \\ 3 & \text{if } \tau_2 < d_{12} \leq \tau_3 \\ 4 & \text{if } \tau_3 < d_{12} < +\infty \end{cases}$$

- The same Brown's Excel macro can be used because Mplus automatically detect the number of categories. For the tirt package, simply change the family in the make_TIRT_data as cumulative.

TIRT : Graded Preference

- Blocked Graded preference have multiple graded preference using the same statements:
- It is possible that the blocked graded preference is not consistent

	++	+	+	++	
1. Once I begin a project I finish it at all costs	X				2. I usually do what others recommend me to do
1. Once I begin a project I finish it at all costs				X	3. I often anticipate the needs of other
2. I usually do what others recommend me to do	X				3. I often anticipate the needs of other

From this example, $1 > 2$, $2 > 3$, but $1 < 3$

TIRT : Graded Preference

- Brown & Maydeu-Olivares (2018) proposed to add intrasensitivity error terms in some ds. For example, in the block of 3,

$$d_{12} = t_1 - t_2 + r_{12}$$

$$d_{13} = t_1 - t_3 + r_{13}$$

$$d_{23} = t_2 - t_3 + r_{23}$$

- However, the tirt package cannot handle intrasensitivity error. The best shot is to adjust Brown's Excel macro results.

TIRT : Proportional Preference

- Because the difference between proportion can be viewed as continuous variable, the differences between proportions can be handled as observed ds .

	Points (100 in total)
1. I pay attention to details	33
2. I have a good word for everyone	67

$$d_{12} = 33 - 67 = -34$$

TIRT : Proportional Preference

- For proportional preference with more than two statements, the intrasensitivity error is not needed because the inconsistency would not exist:

	Points (100 in total)
1. I pay attention to details	50
2. I change my mood a lot	35
3. I have a good word for everyone	10
4. I catch on to things quickly	5

$$d_{12} = 50 - 35 = 15$$

$$d_{13} = 50 - 10 = 40$$

$$d_{14} = 50 - 5 = 45$$

$$d_{23} = 35 - 10 = 25$$

$$d_{24} = 35 - 5 = 30$$

$$d_{34} = 10 - 5 = 5$$

- Brown's Excel macro is not recommended because you need to change item type to continuous so all thresholds should be deleted and fix parameters should be adjusted. For the tirt package, For the tirt package, simply change the family in the make_TIRT_data as gaussian.

How to design a test with ranking items

- Ranking items always provide tremendously less information than single-stimulus items. Don't mess with it if you don't need to.
- Unequally keyed items are good because it reduces ipsativity and the person scores are more reliable. However, it is fakable.
- Thus, to prevent faking, you probably need equally keyed items.

How to design a test with ranking items

- Bürkner (2022) analytically and numerically tested TIRT under many conditions and found that, to get person reliability $> .8$, the test with equally keyed items should have following properties:
 - At least 90 item pairs
 - At least five response categories (can retain 65% of information compared to continuous [binary retains only 13%])
 - At least ten traits are measured. Developers probably need untargeted constructs.
 - Some traits are negatively correlated where all traits are positive when they are socially desirable.
 - At least average factor loadings of .8
 - At least factor loading difference of .2
- If some are more restrictive, others can be relaxed.

How to design a test with ranking items

Thus, the development procedure of graded preference should be as follows:

- Select 10 or more traits
- Write more than 180, 60, and 30 items for block of 2, 3, and 4, respectively
- Run single-stimulus questionnaire
- Survey social desirability of each item
- Select items with high and medium factor loadings without cross loadings

How to design a test with ranking items

- Group items so that in each block
 - Either all items are positive (socially desirable), or all items are negative (socially undesirable)
 - Factor loadings should be all positive and the magnitude are varied.
 - All items should have similar social desirability averages.
- Grouping can be done by the autofc package

How to design a test with ranking items

- Write the graded preference test
- Collect the data of graded preference test to obtain empirical reliability and potentially collect additional variables to get validity evidence
- Norming
- Writing test manual

Miscellaneous

- Multidimensional scaling is similar to TIRT but the factor (dimension) are exploratory.