Cluster Analysis

using Latent Categorical Analysis and Factor Mixture Analysis

Intro

Today's goal:

Teach how to do cluster analysis in Mplus

Outline:

- Explain the idea behind cluster analysis
- Latent Categorical Analysis (LCA)
- Factor Mixture Analysis (FMA)

Cluster Analysis

Why do it?

Cluster Analysis

Putting people into distinct groups…

…based on how they answer certain questions

…based on behavioral patterns

…etc

Two versions:

Based on “raw data”: Latent Categorical Analysis

Based on factors: Factor Mixture Analysis

Dataset

ID	Items
1	Wall
2	Status updates
3	Shared links
4	Notes
5	Photos
6	Hometown
7	Location (city)
8	Location (state/province)
9	Residence (street address)
10	Employer
11	Phone number
12	Email address
13	Religious views
14	Interests (favorite movies, etc.)
15	Facebook groups
16	Friend list

Background

Information disclosure behavior research: two approaches

1. Each item is a separate decision
   - No assumptions about correlations
   - No overall measure of disclosure tendency
   - No explanation of how behaviors come about
   - No suggestion how they can be influenced

Verdict: not very useful

Background

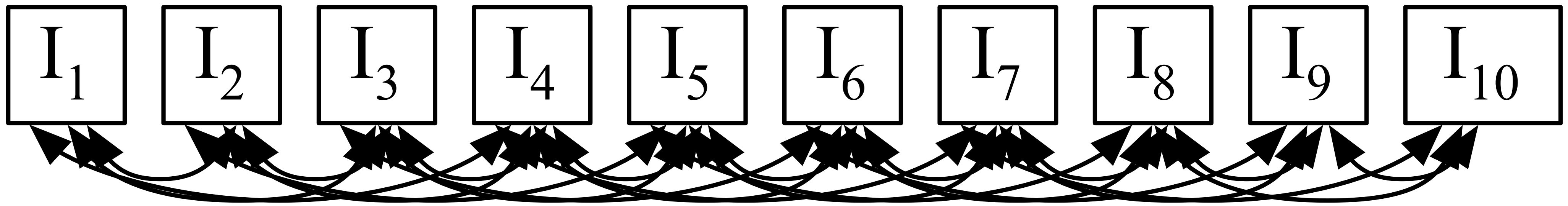
Information disclosure behavior research: two approaches

2. Aggregate of decisions is a single scale
  - Sums individual disclosures to get a "score"
  - Enables researchers to find antecedents
  - Implicit assumption of unidimensionality
  - Implicit assumption of exchangeability

Verdict: might oversimplify the structure of the behavior

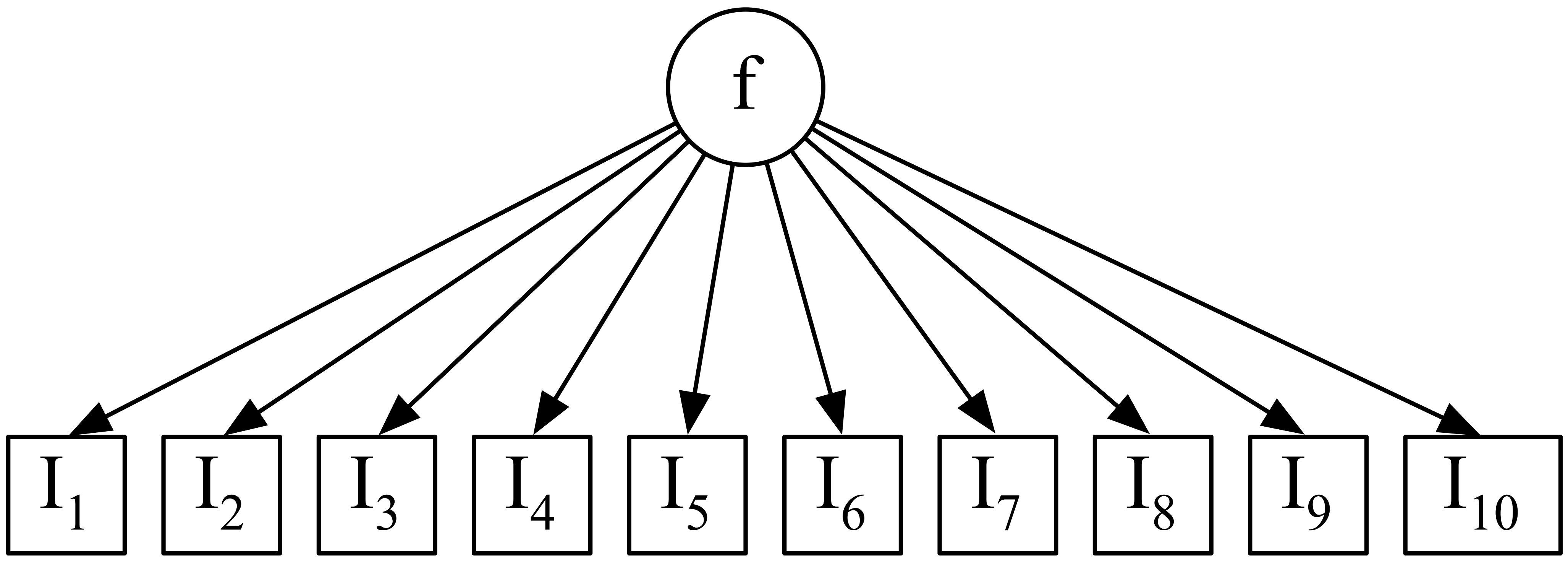
Hypotheses

Disclosures are correlated:



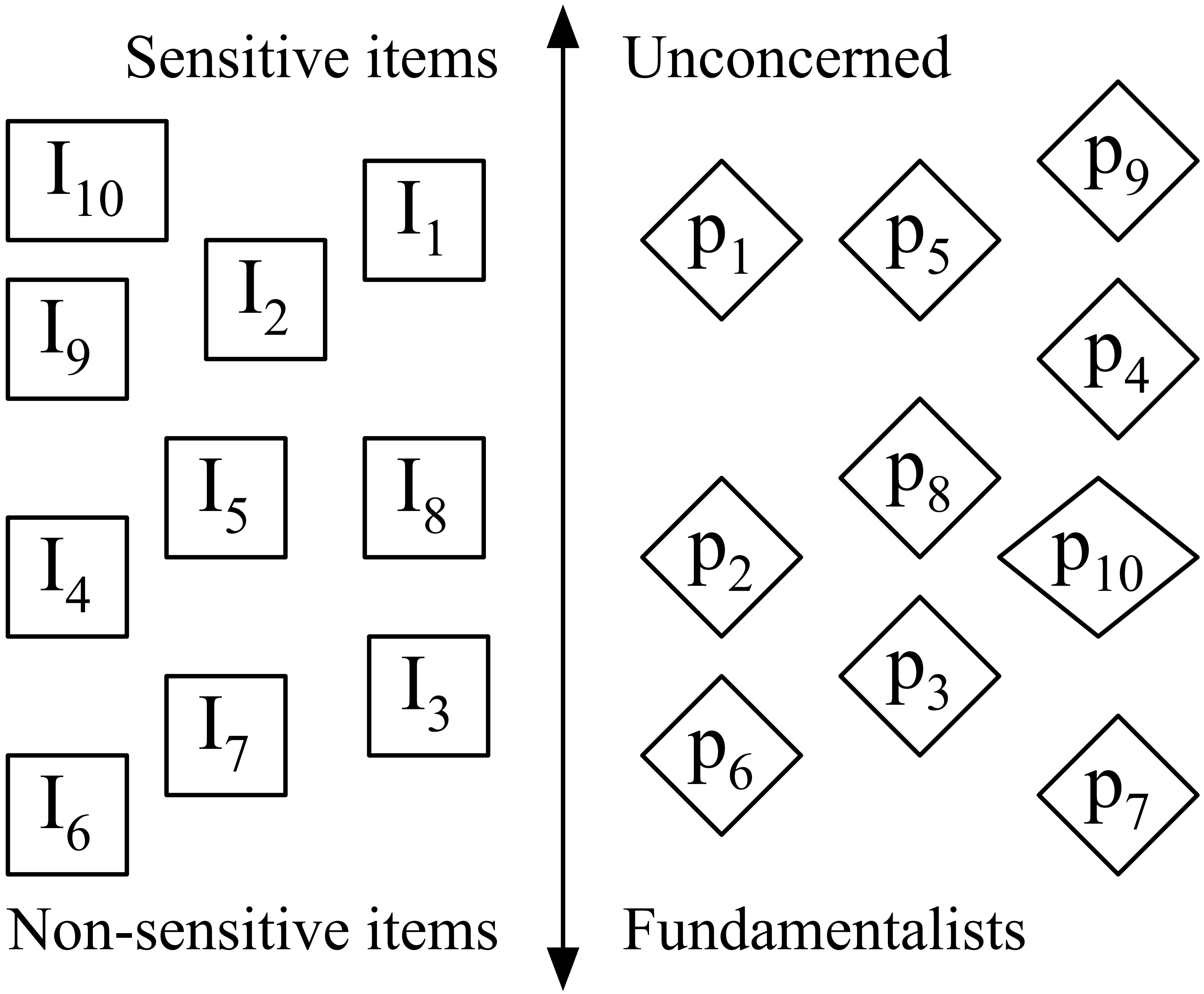
Hypotheses

Disclosures are unidimensional:



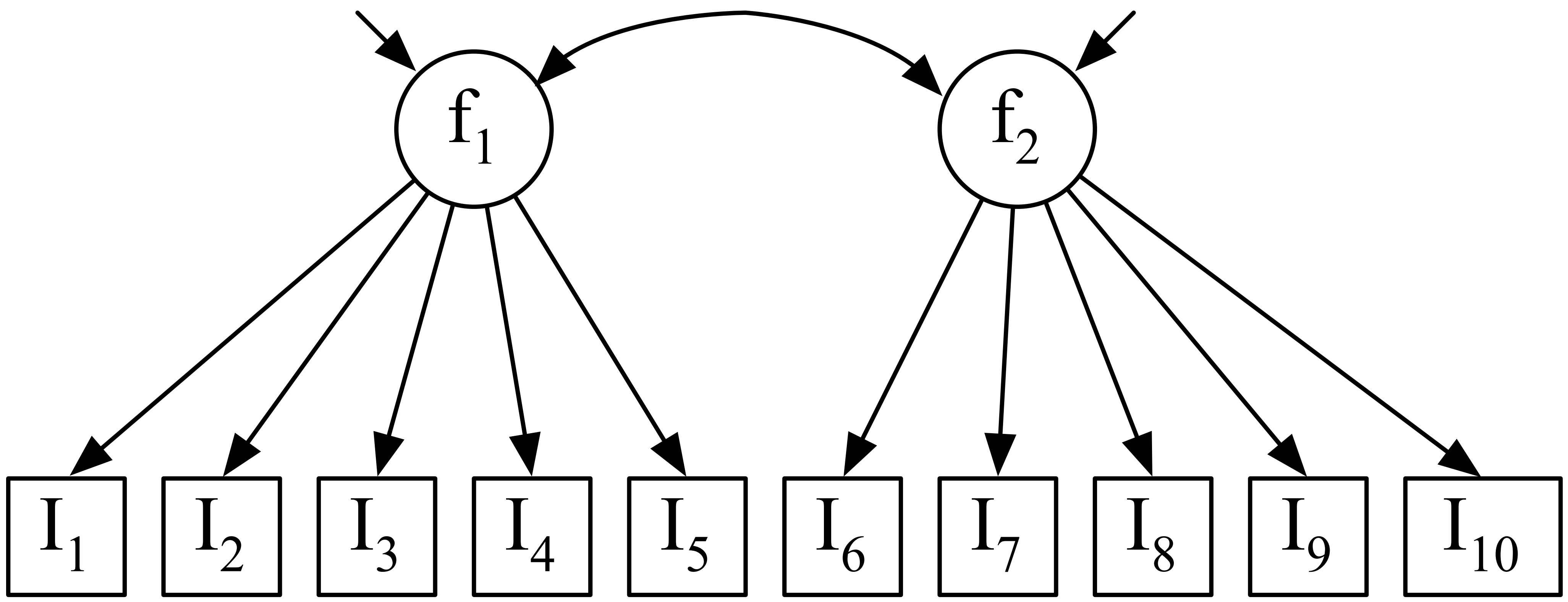
Hypotheses

Disclosures are unidimensional:



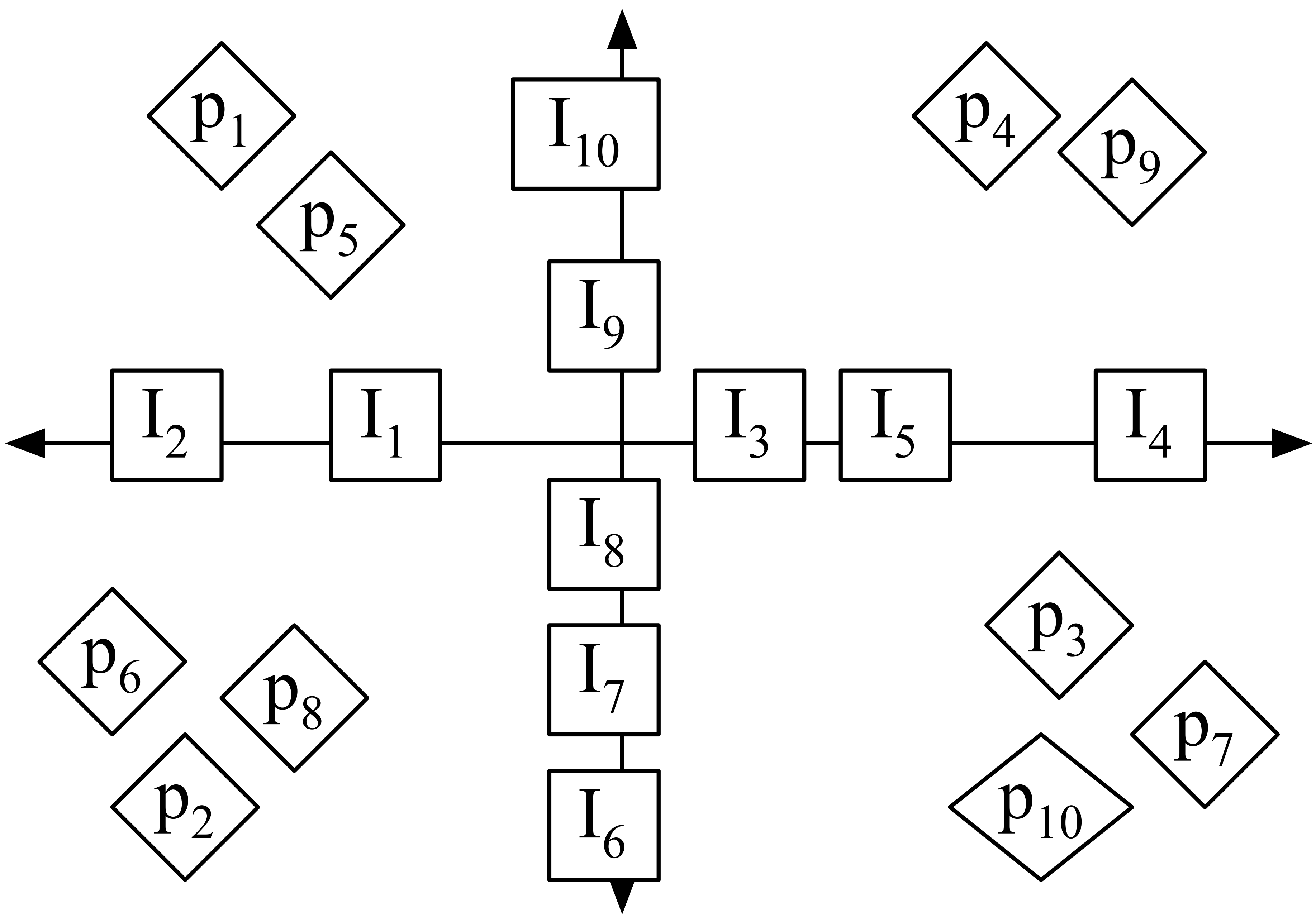
Hypotheses

Disclosures are multidimensional:



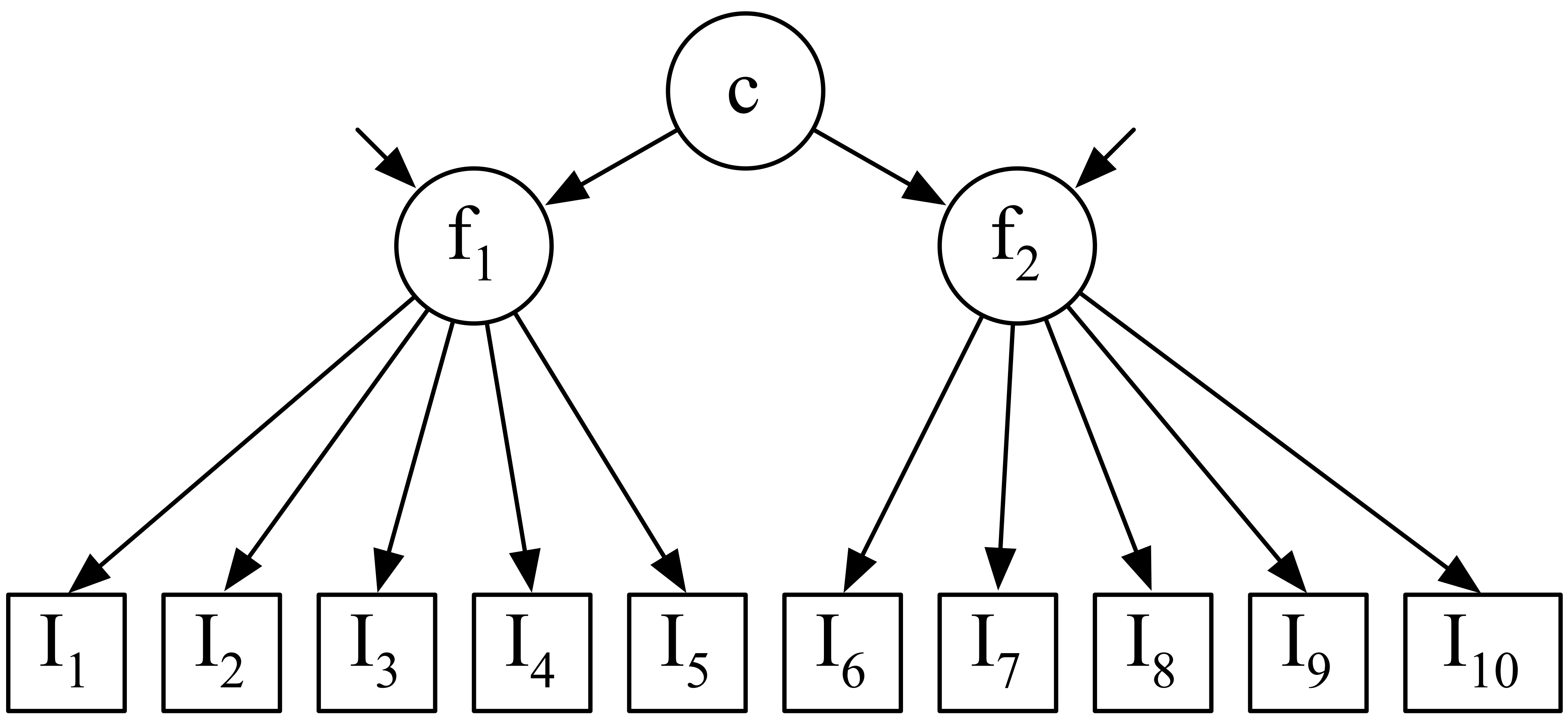
Hypotheses

Disclosures are multidimensional:



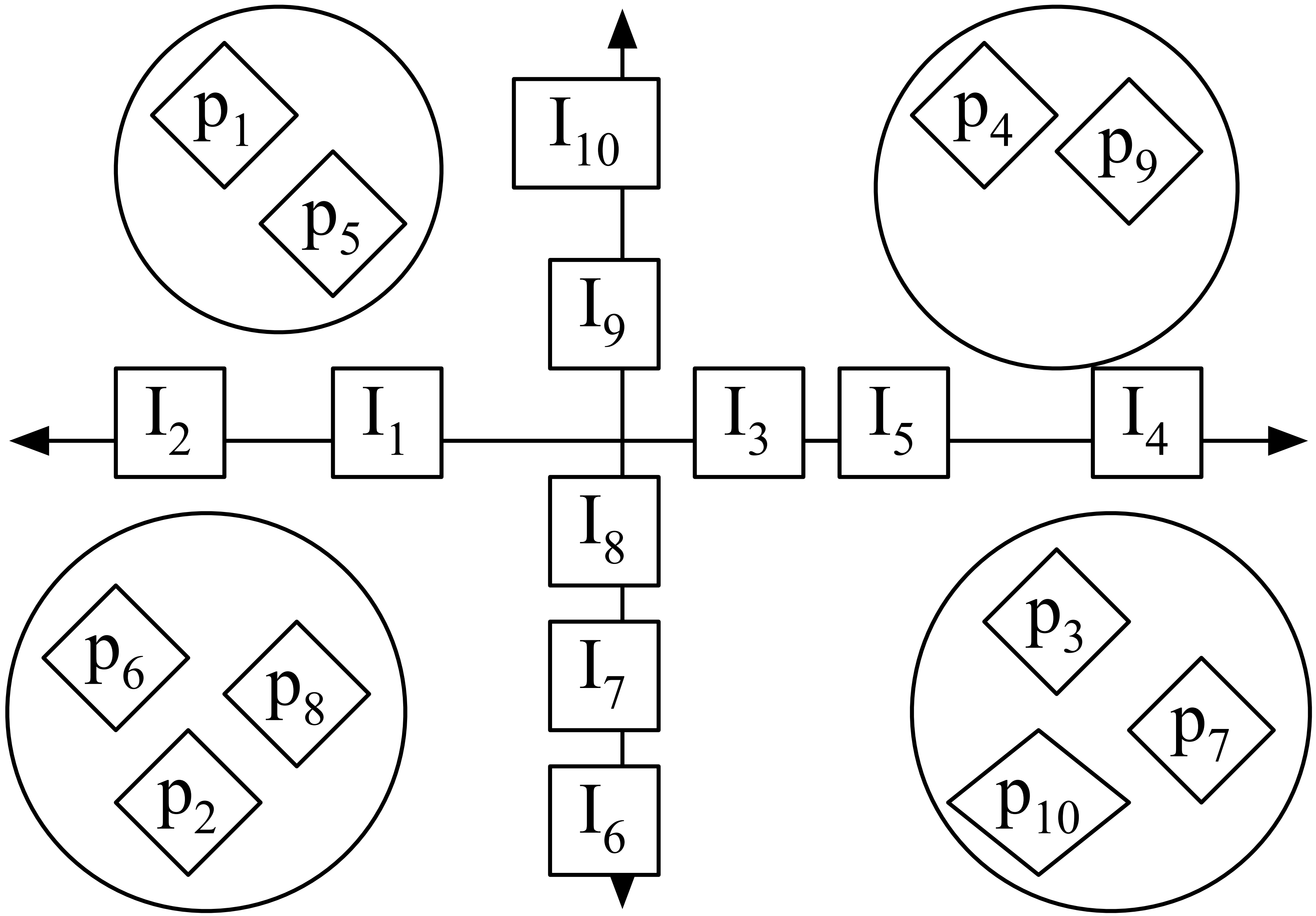
Hypotheses

People can be classified on these dimensions:



Hypotheses

People can be classified on these dimensions:



Hypotheses

Information disclosure behaviors are multidimensional

- Different people have different tendencies to disclose different types of information
- Not one "disclosure tendency", but several!

There exist distinct groups of people with different "disclosure profiles"

- E.g., one group does not disclose location items, while another group does not disclose opinion items

Hypotheses

Privacy groups, that sounds familiar...

- Privacy fundamentalists, pragmatists, and unconcerned (Westin et al., 1981; Harris et al., 2003)

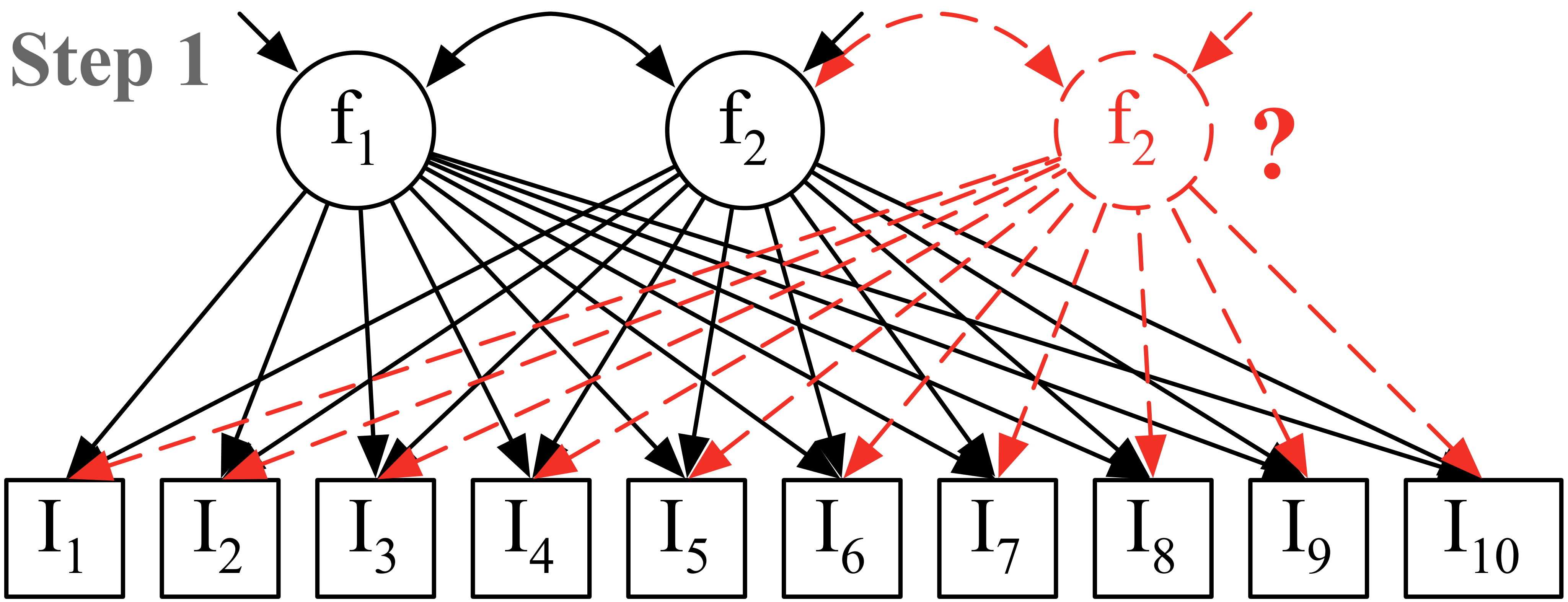
Ours is different:

- Based on behavior rather than attitudes
- Not just a difference in degree, but a difference in kind

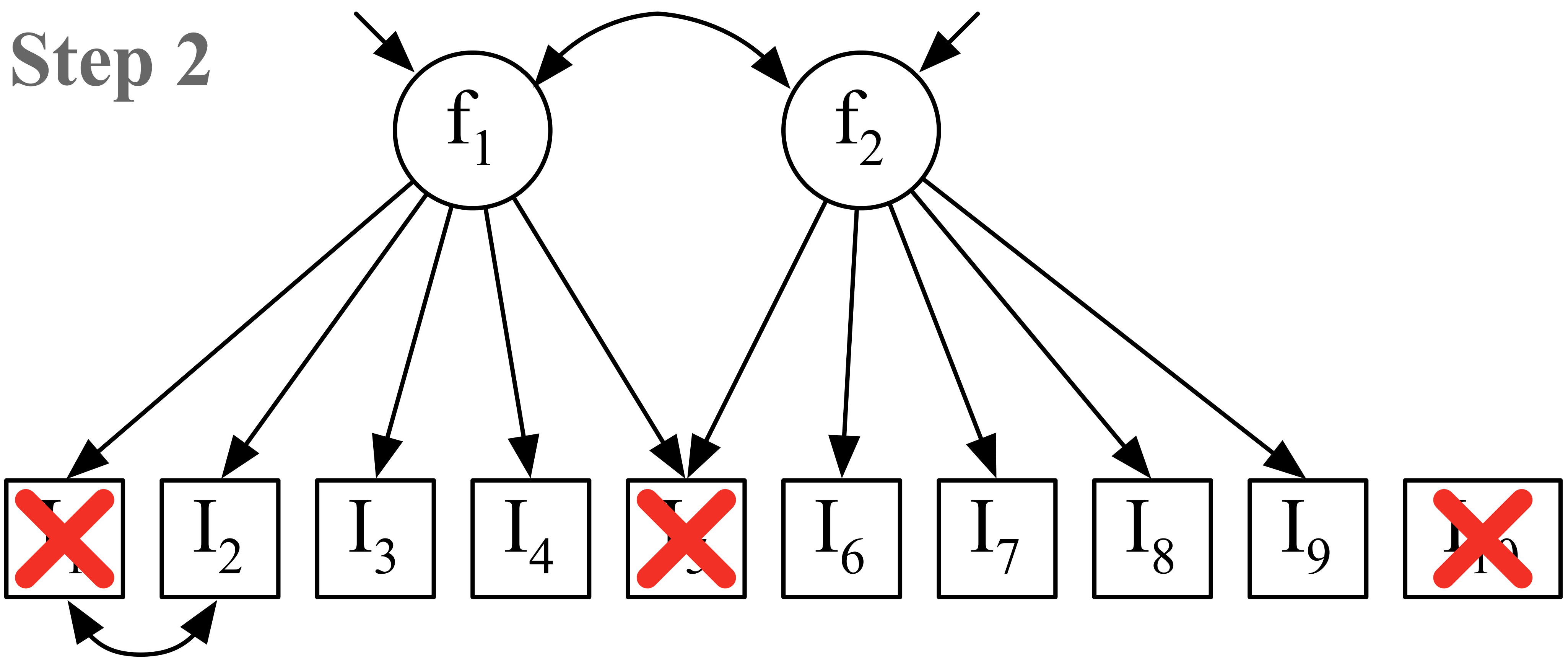
Procedure

ID	Items
1	Wall
2	Status updates
3	Shared links
4	Notes
5	Photos
6	Hometown
7	Location (city)
8	Location (state/province)
9	Residence (street address)
10	Employer
11	Phone number
12	Email address
13	Religious views
14	Interests (favorite movies, etc.)
15	Facebook groups
16	Friend list

Step 1+2



Step 1+2



Initial model (EFA)

	Factor1	Factor2	Factor3	Factor4
cwall	0.810			
cstatus	0.942			
clinks	0.776		0.146	
cnotes	0.790			0.125
cphoto	0.569	0.209		0.140
ctown	0.145	0.698	0.116	
cloccity		0.976		
clocstate		0.960		
clocadress		0.111	-0.105	0.746
cemployer	-0.156	0.311	0.297	0.403
cphone				0.934
cemail			0.211	0.648
creligious			0.810	
cinterest			0.858	
cgroups	0.138		0.755	
cfriends	0.306	0.112	0.462	

Final factors (CFA)

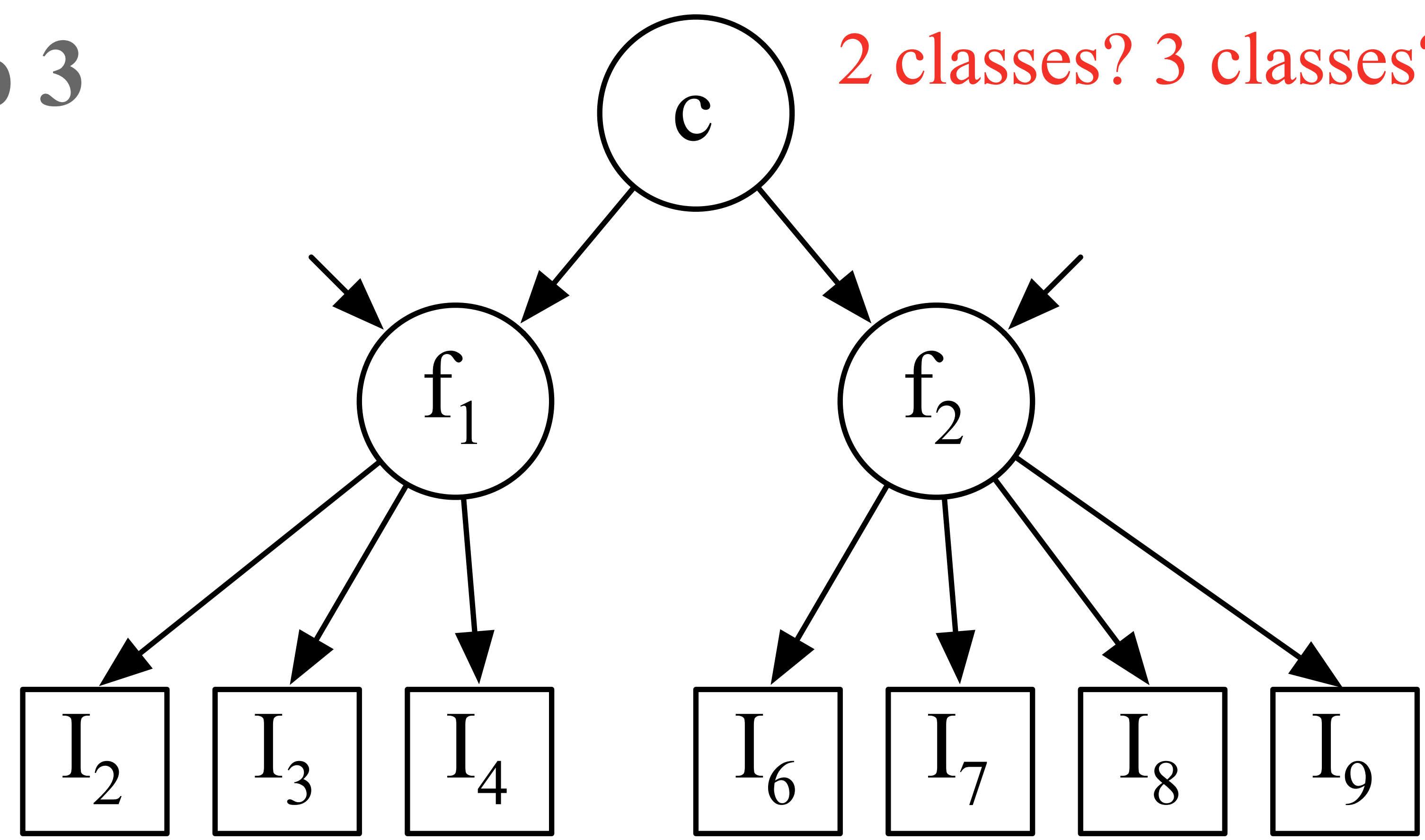
Type of data	ID	Items
Facebook activity	1	Wall
	2	Status updates
	3	Shared links
	4	Notes
	5	Photos
Location	6	Hometown
	7	Location (city)
	8	Location (state/province)
Contact info	9	Residence (street address)
	11	Phone number
	12	Email address
Life/interests	13	Religious views
	14	Interests (favorite movies, etc.)
	15	Facebook groups

Step 3

Factor Mixture Analysis!

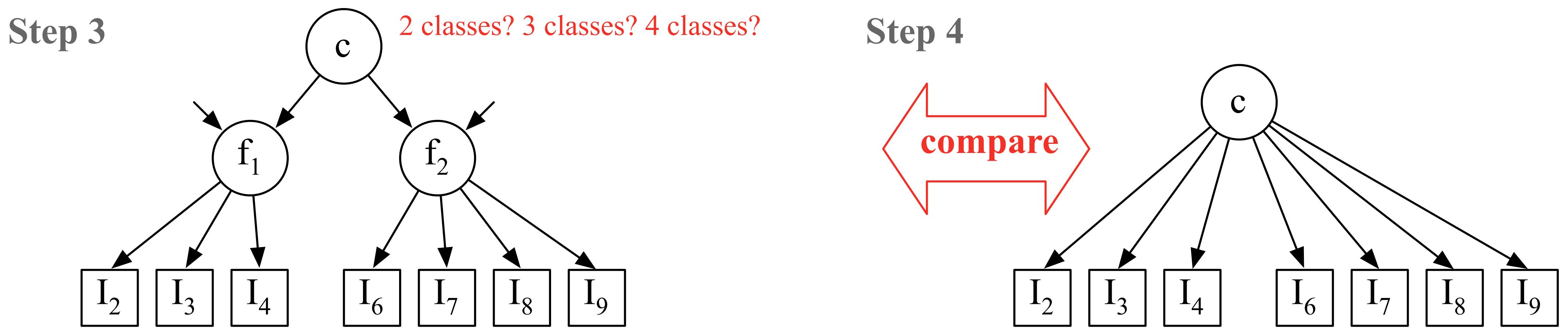
Step 3

2 classes? 3 classes? 4 classes?



Step 4

Factor Mixture Analysis!



Latent Categorical Analysis!

Distinction

LCA: cluster people on the value of the items

Does not assume a latent factor structure

FMA: cluster people on the value of the factors

Assumes a latent factor structure

Sometimes they show essentially the same result

But not always!

LCA in Mplus

How to conduct Latent Categorical Analysis

LCA

Under VARIABLE:

Specify the number of classes: classes = c(2)

Under ANALYSIS:

Specify mixture model: type = mixture

Optionally, specify iterations etc

LCA

```
DATA: file = fdatam.csv;

variable:
    names are
        cwall cstatus clinks cnotes cphoto ctown
        cloccity clocstate clocadress cemployer
        cphone cemail creligious
        cinterest cgroups cfriends
    ;

    usev are
        cwall cstatus clinks cnotes cphoto ctown
        cloccity clocstate clocadress
        cphone cemail creligious
        cinterest cgroups
    ;

    classes = c(2);

analysis:
    type = mixture;
```

Process

Model is going to run with two random clusters

Algorithm adjusts values to create maximum separation between clusters

- 10 initial iterations, plus 4 final optimization steps

Once done, the model restarts with two new random clusters

- 20 random starts

The best results are reported

Results (c=2)

```
RANDOM STARTS RESULTS RANKED FROM THE BEST TO THE WORST LOGLIKELIHOOD VALUES

Final stage loglikelihood values at local maxima, seeds, and initial stage
start numbers:

          -9310.519  637345           19
          -9310.519  573096           20
          -9310.519  285380           1
          -9310.519  195873           6



THE BEST LOGLIKELIHOOD VALUE HAS BEEN REPLICATED.  RERUN WITH AT LEAST TWICE
THE
RANDOM STARTS TO CHECK THAT THE BEST LOGLIKELIHOOD IS STILL OBTAINED AND
REPLICATED.
```

Process

Is the final result we found the best possible result?

It was replicated in 4/20 random starts

Let's run with 200 starts, and check again!

Also, let's increase the number of initial iterations to 20, and the number of final optimizations to 10

Code:

```
starts = 200 10;

sitter = 20;
```

Results (c=2)

```
RANDOM STARTS RESULTS RANKED FROM THE BEST TO THE WORST LOGLIKELIHOOD VALUES

Final stage loglikelihood values at local maxima, seeds, and initial stage
start numbers:

          -9310.519  417035           149
          -9310.519  754100           56
          -9310.519  496881           192
          -9310.519  407168           44
          -9310.519  475420           71
          -9310.519  950604           172
          -9310.519  963053           43
          -9310.519  207896           25
          -9310.519  830392           35
          -9310.519  846194           93


THE BEST LOGLIKELIHOOD VALUE HAS BEEN REPLICATED.  RERUN WITH AT LEAST TWICE
THE
RANDOM STARTS TO CHECK THAT THE BEST LOGLIKELIHOOD IS STILL OBTAINED AND
REPLICATED.
```

More results (c=2)

```
MODEL FIT INFORMATION

Number of Free Parameters                        43

Loglikelihood

          H0 Value                        -9310.519
          H0 Scaling Correction Factor       1.1612
            for MLR

Information Criteria

          Akaike (AIC)                    18707.038
          Bayesian (BIC)                  18874.020
          Sample-Size Adjusted BIC        18737.603
            (n* = (n + 2) / 24)
```

More results (c=2)

```
FINAL CLASS COUNTS AND PROPORTIONS FOR THE LATENT CLASSES
BASED ON THEIR MOST LIKELY LATENT CLASS MEMBERSHIP

Class Counts and Proportions

    Latent
   Classes

       1              202          0.56267
       2              157          0.43733


CLASSIFICATION QUALITY

     Entropy                          0.951
```

More results (c=2)

	Estimate	S.E.	Est./S.E.	Two-Tailed P-Value
Latent Class 1				
Means				
CWALL	2.544	0.150	16.973	0.000
CSTATUS	2.174	0.130	16.749	0.000
CLINKS	2.664	0.139	19.101	0.000
CNOTES	1.943	0.108	18.006	0.000
CPHOTO	1.682	0.099	16.919	0.000
CTOWN	2.731	0.125	21.922	0.000
CLOCCITY	2.565	0.125	20.563	0.000
CLOCSTATE	2.818	0.131	21.429	0.000
CLOCADRESS	1.184	0.040	29.384	0.000
CPHONE	1.077	0.023	46.952	0.000
CEMAIL	1.665	0.083	19.988	0.000
CRELIGIOUS	3.565	0.143	24.942	0.000
CINTEREST	3.635	0.136	26.781	0.000
CGROUPS	3.366	0.132	25.418	0.000

More results (c=2)

	Estimate	S.E.	Est./S.E.	Two-Tailed P-Value
Latent Class 2				
Means				
CWALL	5.430	0.125	43.485	0.000
CSTATUS	5.527	0.119	46.282	0.000
CLINKS	5.492	0.122	44.990	0.000
CNOTES	4.992	0.150	33.210	0.000
CPHOTO	4.742	0.167	28.447	0.000
CTOWN	5.439	0.143	38.040	0.000
CLOCCITY	5.029	0.173	29.127	0.000
CLOCSTATE	5.246	0.162	32.480	0.000
CLOCADRESS	2.919	0.184	15.841	0.000
CPHONE	2.605	0.169	15.416	0.000
CEMAIL	3.757	0.181	20.711	0.000
CRELIGIOUS	5.117	0.133	38.471	0.000
CINTEREST	5.598	0.115	48.888	0.000
CGROUPS	5.643	0.111	50.925	0.000

Results

Two classes: one low, one high

What about the 3-class solution?

- Change **classes = c(3)**;
- To compare against 2 classes, add **output: tech11;**
- Long wait? Add **processors = 4;** (or 8) to make things parallel!

Results (c=3)

```
RANDOM STARTS RESULTS RANKED FROM THE BEST TO THE WORST LOGLIKELIHOOD VALUES

Final stage loglikelihood values at local maxima, seeds, and initial stage
start numbers:

           -8980.584  761633           50
           -8980.584  414284           158
           -8980.584  860772           174
           -8980.584  544048           87
           -8980.584  479273           156
           -8980.584  576596           99
           -8980.584  804561           59
           -8980.584  286735           175
           -8980.584  458181           189
           -8980.584  939709           112


THE BEST LOGLIKELIHOOD VALUE HAS BEEN REPLICATED.  RERUN WITH AT LEAST TWICE
THE
RANDOM STARTS TO CHECK THAT THE BEST LOGLIKELIHOOD IS STILL OBTAINED AND
REPLICATED.
```

More results (c=3)

```
MODEL FIT INFORMATION

Number of Free Parameters                              58

Loglikelihood

          H0 Value                              -8980.584
          H0 Scaling Correction Factor             1.3522
            for MLR

Information Criteria

          Akaike (AIC)                          18077.167
          Bayesian (BIC)                        18302.400 (vs 18874.020)
          Sample-Size Adjusted BIC              18118.395
            (n* = (n + 2) / 24)
```

More results (c=3)

```
FINAL CLASS COUNTS AND PROPORTIONS FOR THE LATENT CLASSES
BASED ON THEIR MOST LIKELY LATENT CLASS MEMBERSHIP

Class Counts and Proportions

    Latent
   Classes

       1              164          0.45682
       2              130          0.36212
       3               65          0.18106


CLASSIFICATION QUALITY

     Entropy                            0.957 (vs 0.951)
```

More results (c=3)

	Estimate	S.E.	Est./S.E.	Two-Tailed P-Value
Latent Class 1				
Means				
CWALL	2.258	0.142	15.914	0.000
CSTATUS	1.912	0.104	18.407	0.000
CLINKS	2.354	0.126	18.729	0.000
CNOTES	1.666	0.094	17.686	0.000
CPHOTO	1.443	0.082	17.694	0.000
CTOWN	2.504	0.170	14.687	0.000
CLOCCITY	2.329	0.181	12.865	0.000
CLOCSTATE	2.554	0.189	13.534	0.000
CLOCADRESS	1.158	0.049	23.444	0.000
CPHONE	1.057	0.021	51.179	0.000
CEMAIL	1.580	0.086	18.291	0.000
CRELIGIOUS	3.263	0.169	19.271	0.000
CINTEREST	3.251	0.174	18.735	0.000
CGROUPS	3.002	0.156	19.236	0.000

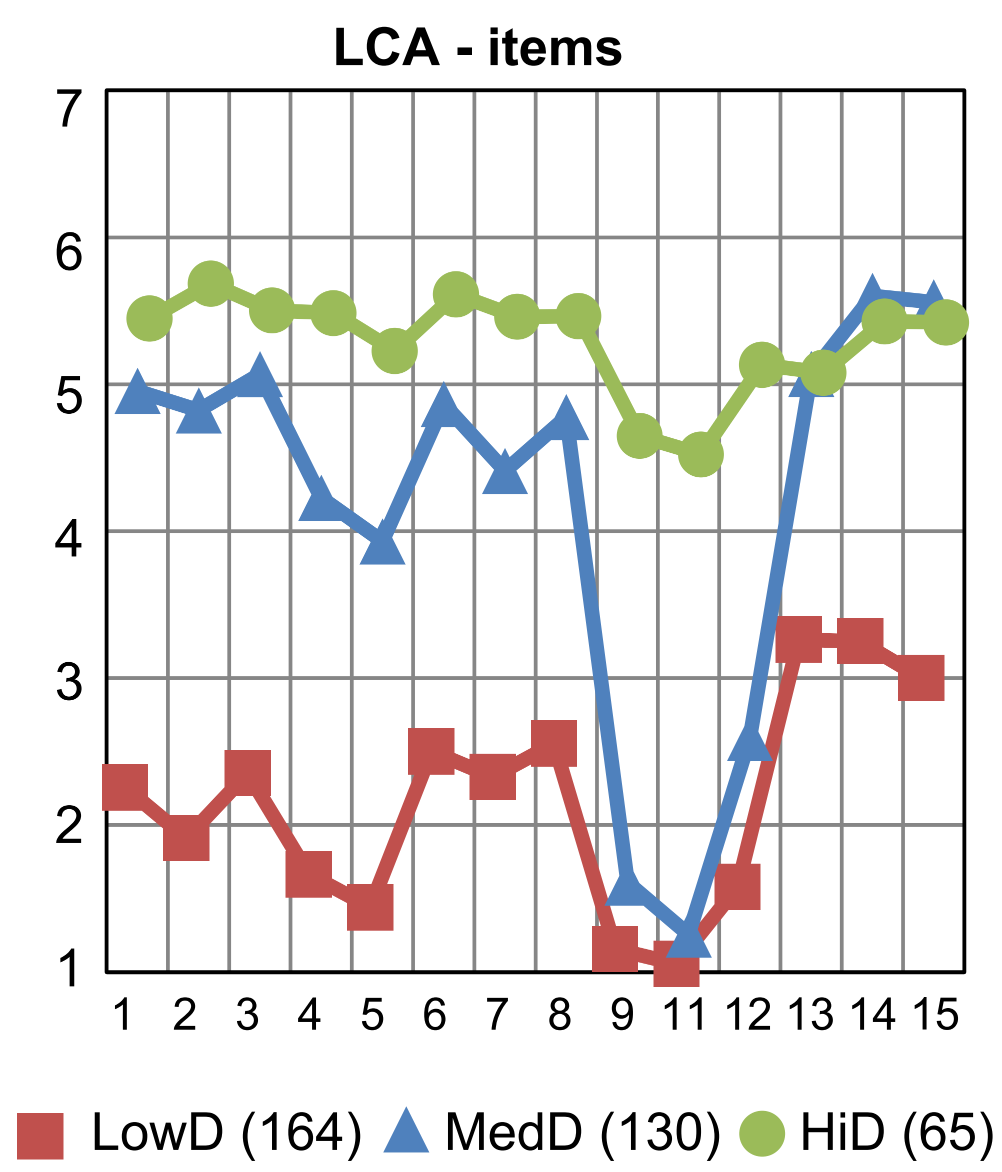
More results (c=3)

	Estimate	S.E.	Est./S.E.	Two-Tailed P-Value
Latent Class 2				
Means				
CWALL	4.956	0.227	21.812	0.000
CSTATUS	4.822	0.234	20.590	0.000
CLINKS	5.069	0.195	26.048	0.000
CNOTES	4.228	0.206	20.490	0.000
CPHOTO	3.931	0.217	18.133	0.000
CTOWN	4.866	0.176	27.576	0.000
CLOCCITY	4.410	0.184	23.958	0.000
CLOCSTATE	4.777	0.177	26.964	0.000
CLOCADRESS	**1.610**	**0.112**	**14.410**	**0.000**
CPHONE	**1.256**	**0.061**	**20.544**	**0.000**
CEMAIL	**2.593**	**0.169**	**15.306**	**0.000**
CRELIGIOUS	5.071	0.154	32.849	0.000
CINTEREST	5.602	0.123	45.488	0.000
CGROUPS	5.558	0.149	37.411	0.000

More results (c=3)

	Estimate	S.E.	Est./S.E.	Two-Tailed P-Value
Latent Class 3				
Means				
CWALL	5.448	0.192	28.360	0.000
CSTATUS	5.685	0.153	37.132	0.000
CLINKS	5.503	0.169	32.637	0.000
CNOTES	5.485	0.171	32.039	0.000
CPHOTO	5.227	0.195	26.799	0.000
CTOWN	5.612	0.174	32.202	0.000
CLOCCITY	5.460	0.171	31.937	0.000
CLOCSTATE	5.465	0.181	30.173	0.000
CLOCADRESS	4.649	0.280	16.609	0.000
CPHONE	4.523	0.200	22.666	0.000
CEMAIL	5.133	0.154	33.375	0.000
CRELIGIOUS	5.079	0.192	26.492	0.000
CINTEREST	5.428	0.193	28.064	0.000
CGROUPS	5.421	0.147	36.846	0.000

More results (c=3)



More results (c=3)

```
     VUONG-LO-MENDELL-RUBIN LIKELIHOOD RATIO TEST FOR 2 (H0) VERSUS 3
CLASSES

          H0 Loglikelihood Value                              -9310.519
          2 Times the Loglikelihood Difference                  659.870
          Difference in the Number of Parameters                     15
          Mean                                                  186.543
          Standard Deviation                                    211.597
          P-Value                                                0.0326

     LO-MENDELL-RUBIN ADJUSTED LRT TEST

          Value                                                 652.477
          P-Value                                                0.0339
```

Results

More results (c=4)

```
MODEL FIT INFORMATION

Number of Free Parameters                       73

Loglikelihood

          H0 Value                       -8745.883
          H0 Scaling Correction Factor      1.3460
            for MLR

Information Criteria

          Akaike (AIC)                   17637.766
          Bayesian (BIC)                 17921.249 (vs 18302.400)
          Sample-Size Adjusted BIC       17689.657
            (n* = (n + 2) / 24)
```

More results (c=4)

FINAL CLASS COUNTS AND PROPORTIONS FOR THE LATENT CLASSES
BASED ON THEIR MOST LIKELY LATENT CLASS MEMBERSHIP

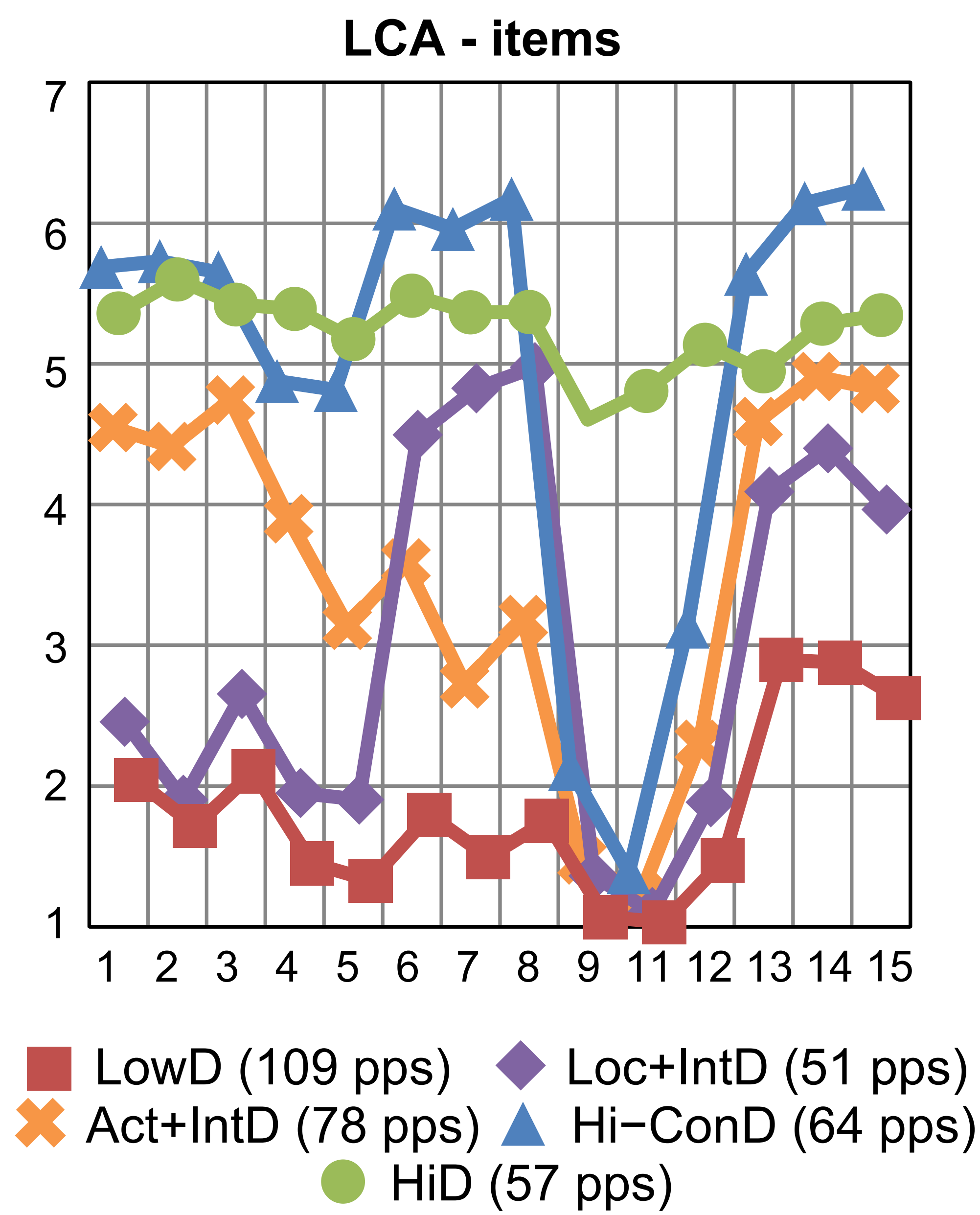
Class Counts and Proportions

Latent Classes		
1	107	0.29805
2	69	0.19220
3	124	0.34540
4	59	0.16435

CLASSIFICATION QUALITY

Entropy **0.929** (vs 0.957)

More results (c=4)



More results (c=4)

```
     VUONG-LO-MENDELL-RUBIN LIKELIHOOD RATIO TEST FOR 3 (H0) VERSUS 4
CLASSES

          H0 Loglikelihood Value                        -8980.584
          2 Times the Loglikelihood Difference            469.401
          Difference in the Number of Parameters               15
          Mean                                             43.297
          Standard Deviation                              229.372
          P-Value                                          0.0316

     LO-MENDELL-RUBIN ADJUSTED LRT TEST

          Value                                           464.142
          P-Value                                          0.0333
```

Results

More results (c=4)

```
MODEL FIT INFORMATION

Number of Free Parameters                              88

Loglikelihood

          H0 Value                              -8607.884
          H0 Scaling Correction Factor             1.5979
            for MLR

Information Criteria

          Akaike (AIC)                          17391.768
          Bayesian (BIC)                        17733.500 (vs 17921.249)
          Sample-Size Adjusted BIC              17454.320
            (n* = (n + 2) / 24)
```

More results (c=4)

```
FINAL CLASS COUNTS AND PROPORTIONS FOR THE LATENT CLASSES
BASED ON THEIR MOST LIKELY LATENT CLASS MEMBERSHIP

Class Counts and Proportions

    Latent
   Classes

       1                78          0.21727
       2               109          0.30362
       3                51          0.14206
       4                57          0.15877
       5                64          0.17827


CLASSIFICATION QUALITY

     Entropy                         0.940 (vs 0.929)
```

More results (c=4)

```
     VUONG-LO-MENDELL-RUBIN LIKELIHOOD RATIO TEST FOR 4 (H0) VERSUS 5
CLASSES

          H0 Loglikelihood Value                        -8745.883
          2 Times the Loglikelihood Difference            275.999
          Difference in the Number of Parameters               15
          Mean                                            733.767
          Standard Deviation                              830.221
          P-Value                                          0.7093

     LO-MENDELL-RUBIN ADJUSTED LRT TEST

          Value                                           272.906
          P-Value                                          0.7106
```

How many classes?

Balance the following criteria

- Minimum of BIC
- Maximum entropy
- Loglikelihood levels off
- p-value of successor > .05 (use Lo-Mendell-Rubin adjusted LRT test, available in output: tech11)
- **Solution makes sense**

FMA in Mplus

How to conduct Factor Mixture Analysis

FMA

Under VARIABLE:

Specify the number of classes: classes = c(2)

Under ANALYSIS:

Specify mixture model: type = mixture

Optionally, specify iterations etc (often needed!)

Under MODEL:

Add %overall% and then the factor model

Prepare to wait :-)

FMA

```
    usev are
        cwall cstatus clinks cnotes cphoto ctown
        cloccity clocstate clocadress
        cphone cemail creligious
        cinterest cgroups
    ;

    classes = c(2);

analysis:
    type = mixture;
    starts = 400 20;
    stiter = 40;
    processors = 8;

model:
    %overall%
    activity BY cwall cstatus clinks cnotes cphoto;
    location BY ctown cloccity clocstate;
    contact  BY clocadress cphone cemail;
    prefs    BY creligious cinterest cgroups;
```

How many classes?

Balance the following criteria

- Minimum of BIC
- Maximum entropy
- Loglikelihood levels off
- p-value of successor > .05 (use Lo-Mendell-Rubin adjusted LRT test, available in output: tech11)
- **Solution makes sense**

Results

Table 9
A comparison of the fit of MFA models with different numbers of classes.

	BIC	Entropy	LL	# of par.	*p*-Value
1 class	16,837		−8277.147	48	
2 classes	16,578	0.973	−8133.179	53	0.0069
3 classes	**16,442**	**0.998**	−8050.552	58	0.0002
4 classes	16,468	0.998	−8048.736	63	**0.407**
5 classes	16,482	0.878	−8041.459	68	0.999
6 classes	**16,351**	0.897	−7960.902	73	0.812
7 classes	16,359	0.852	−7950.412	78	0.893

The bold values are mentioned in the text as indicators of the optimal number of dimensions.

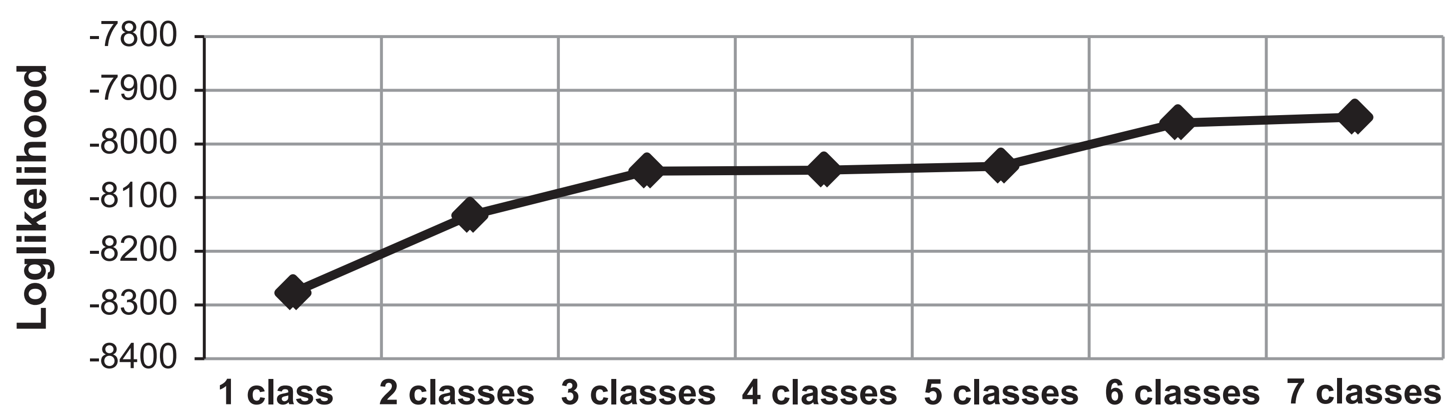
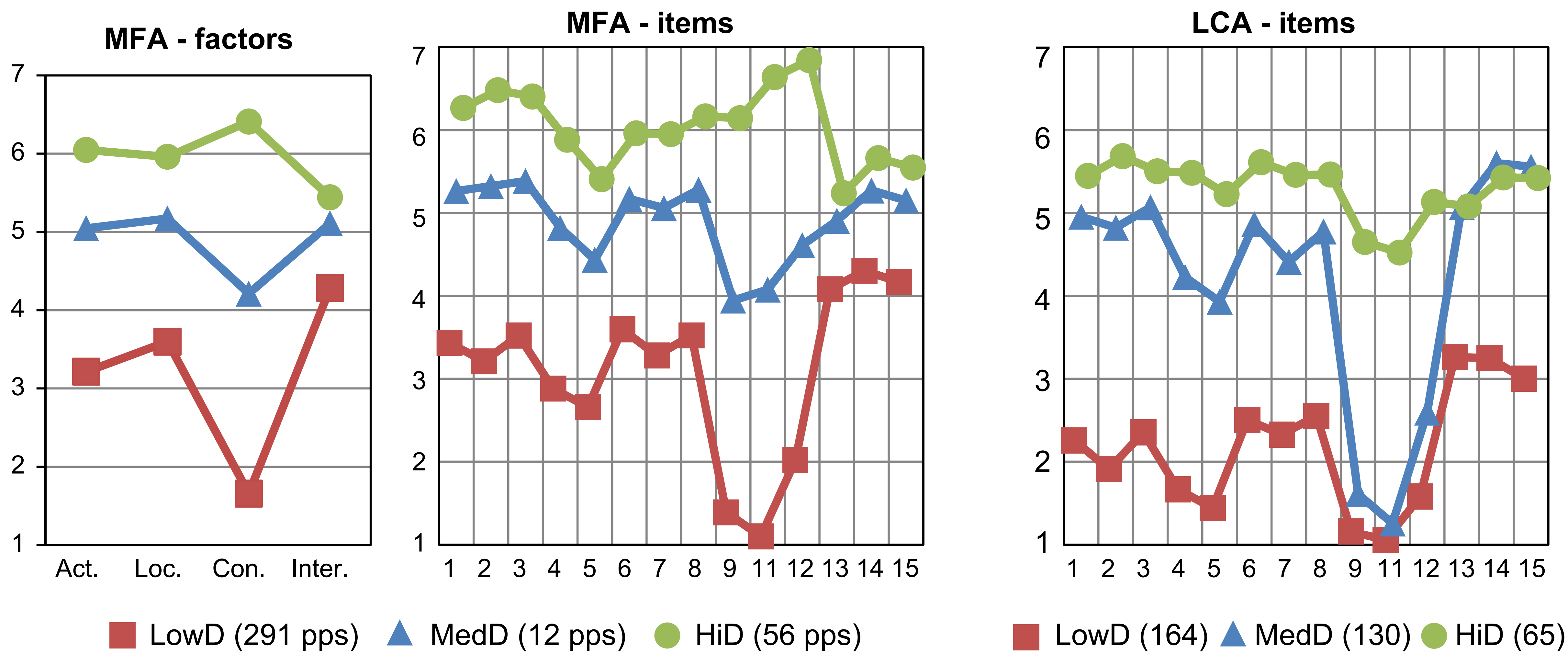
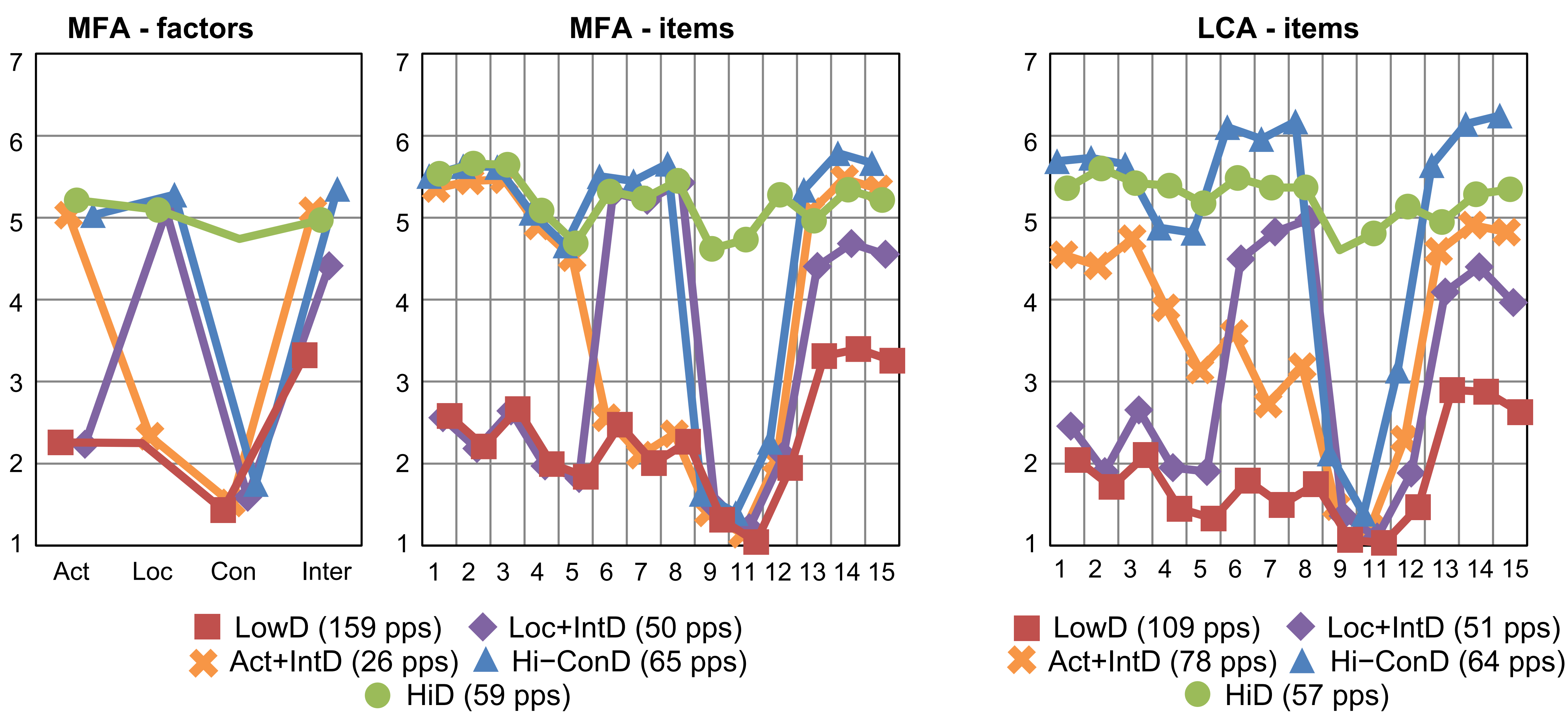


Fig. 8. Change in loglikelihood between subsequent MFA models.

Results


MFA - factors
MFA - items
LCA - items
Act.
Loc.
Con.
Inter.
1 2 3 4 5 6 7 8 9 11 12 13 14 15
LowD (291 pps)
MedD (12 pps)
HiD (56 pps)
LowD (164)
MedD (130)
HiD (65)

Results



MFA - factors
MFA - items
LCA - items
7
6
5
4
3
2
1
Act
Loc
Con
Inter
1 2 3 4 5 6 7 8 9 11 12 13 14 15
LowD (159 pps)
Loc+IntD (50 pps)
Act+IntD (26 pps)
Hi−ConD (65 pps)
HiD (59 pps)
LowD (109 pps)
Loc+IntD (51 pps)
Act+IntD (78 pps)
Hi−ConD (64 pps)
HiD (57 pps)

FMA

Papers:

Knijnenburg et al. (2012): "Dimensionality of information disclosure behavior ", *IJHCS 71 - bit.ly/privdim*

Wisniewski et al. (2016): "Making privacy personal: Profiling social network users to inform privacy education and nudging", *IJHCS 98 - bit.ly/ijhcs2016*

“It is the mark of a truly intelligent person to be moved by statistics.”

T H A N K S !

George Bernard Shaw