Part 4: Advanced

the really cool stuff…

Slides

Feel free to share these slides with anyone

This is version 0.9 (still looking to expand the examples). For the most recent version of these slides, visit www.usabart.nl/QRMS

If you want to use these slides in your own lectures, use the above link for attribution

BREN:ICS INFORMATION AND COMPUTER SCIENCES

Advanced Topics

In this part I discuss the following advanced topics:

- Multi-level regressions and SEM
- Interaction effects in SEM
- Cluster analysis

BREN:ICS INFORMATION AND COMPUTER SCIENCES

Multi-level models

in regression analysis and SEM

Multi-level models

Repeated measurements

e.g. participants make 30 decisions

(Partially) within-subjects design

e.g. participants are randomly assigned to 1 of 3 games, and test it once with sound on and once with sound off

Grouped data

e.g. participants perform tasks in groups of 5

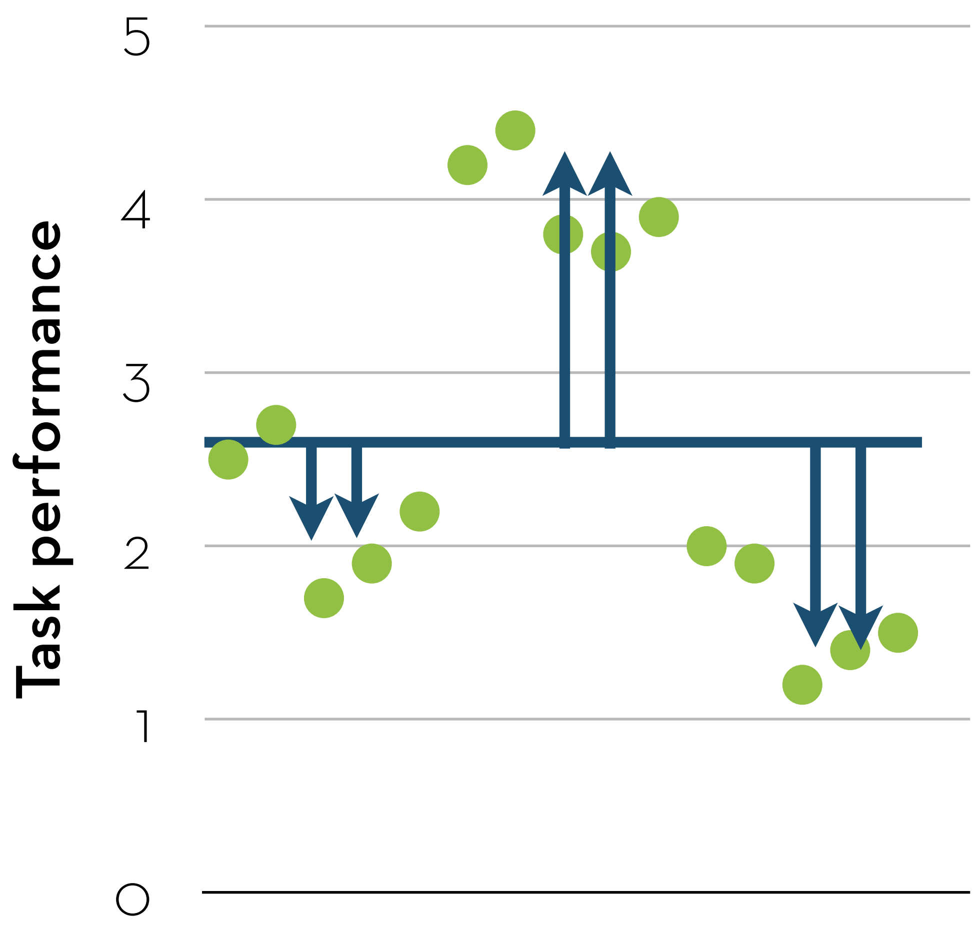
A combination of the above

Correlated errors

Consequence: errors are correlated

- There will be a user-bias (and maybe an task-bias)

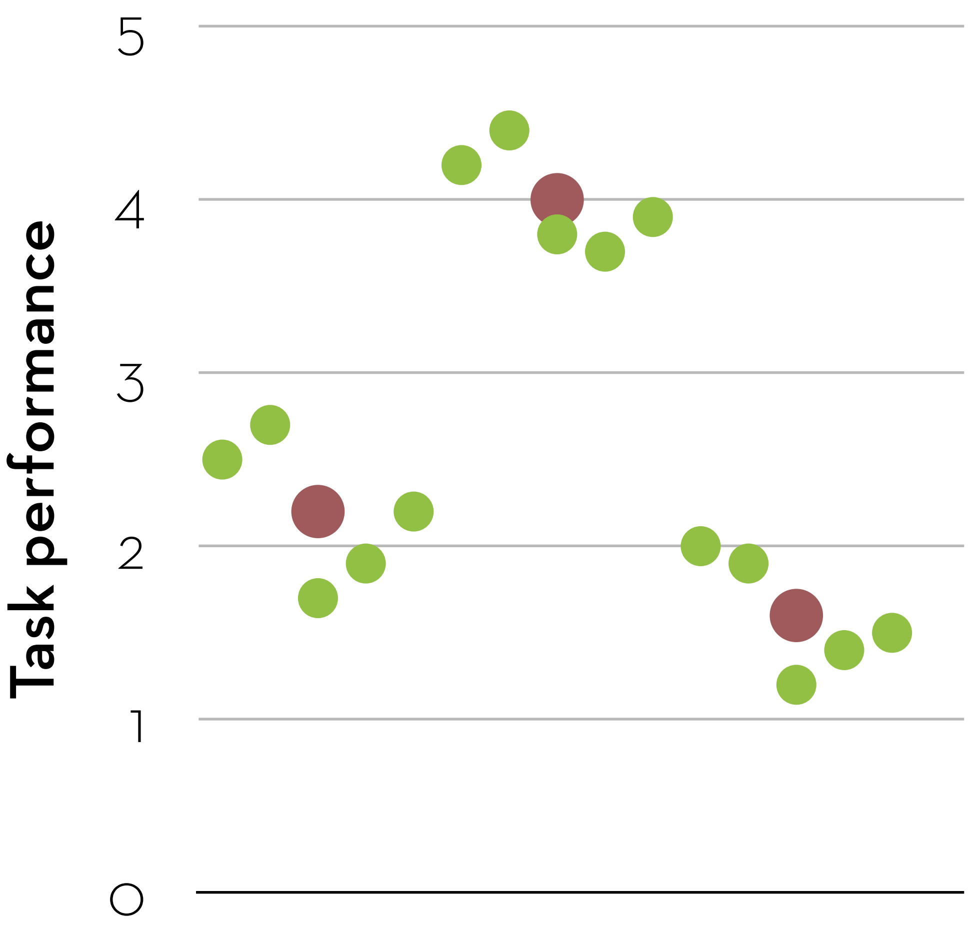
Golden rule: data-points should be **independent**



OK solution...

Take the average of the repeated measurements

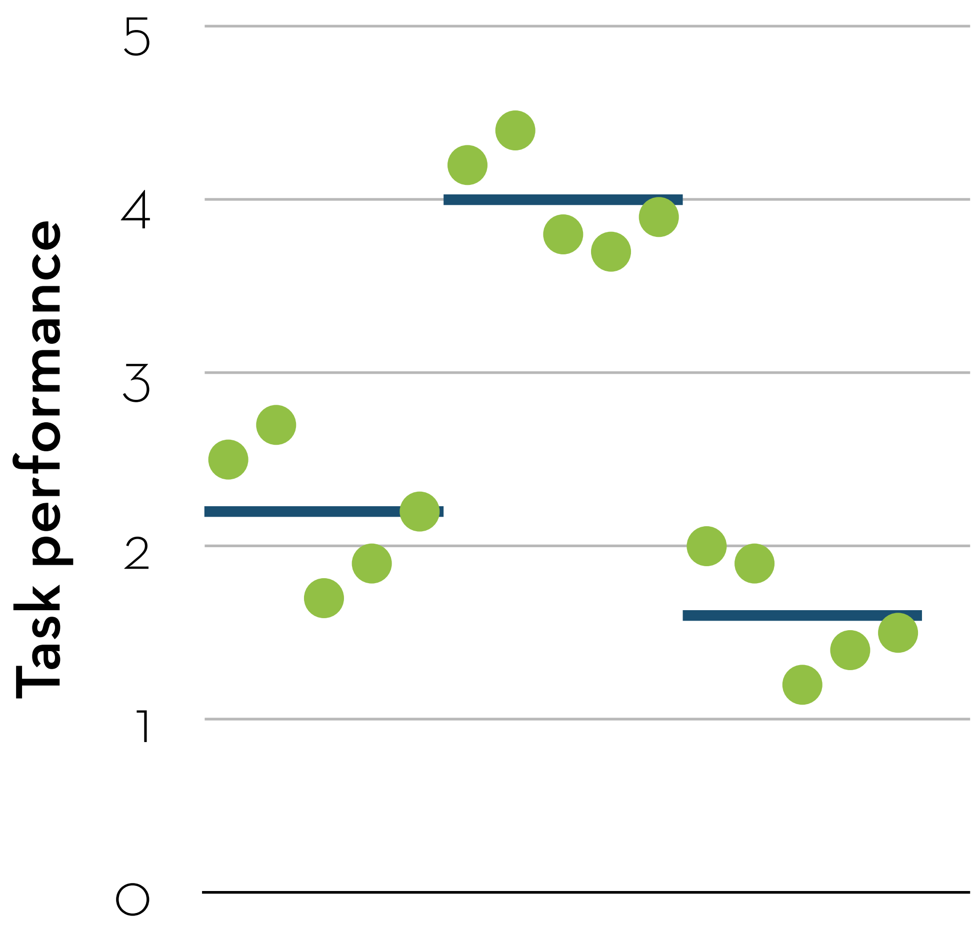
- Reduces the number of observations
- It becomes impossible to make inferences about individual tasks/users/etc.



Good solution

In regression:

- define a random intercept for each user (GLMM)
- impose an error covariance structure (GEE)



INFORMATION AND COMPUTER SCIENCES

Example

Figs here

Example

Data: 396 participants each make 31 disclosure decisions (binary)

Manipulations:

- Between subjects: 5 justification types: 1:none, 2:useful-for-you, 3:%others, 4:useful-for-others, 5:explanation
- Between subjects: request order (counter-balanced)
- Within subjects: questionID (#1-#31)
- Within subjects: percentage (only for justification types 2, 3 and 4)

BREN:ICS INFORMATION AND COMPUTER SCIENCES

Research question

What is the effect of the justification types, and does the percentage displayed in the justification play any role?

Wrong solution

Naive specification in R, using GLM:

```
model1 <- glm(decision ~ fmessage*percentage, family=binomial, data=fat2)
```

Output:

```
Coefficients:
                     Estimate Std. Error z value Pr(>|z|)
(Intercept)          1.585620   0.049354  32.128  < 2e-16 ***
fmessage1           -0.218224   0.067850  -3.216  0.00130 **
fmessage2           -0.514137   0.063370  -8.113 4.93e-16 ***
fmessage3           -0.630636   0.063346  -9.955  < 2e-16 ***
fmessage4           -0.206947   0.067171  -3.081  0.00206 **
percentage          -0.002052   0.001698  -1.209  0.22670
fmessage1:percentage  0.003472   0.002313   1.501  0.13332
fmessage2:percentage  0.006351   0.002176   2.919  0.00351 **
fmessage3:percentage  0.003224   0.002175   1.482  0.13830
fmessage4:percentage  0.002145   0.002317   0.926  0.35457
```

GLMM

GLMM = Generalized Linear Mixed-effects Models

- Works on normal data (LMM) and binary/count data (GLMM)

R package: lme4

- Function: glmer (or lmer)

Better solution

Random intercept for participant (sessionId):

```
model2 <- glmer(decision ~ fmessage*percentage + (1|sessionId),
family=binomial, data=fat2)
```

BREN:ICS INFORMATION AND COMPUTER SCIENCES

Interpretation

The 15283 data points originate from 493 participant

How do we deal with this?

- We could create a separate dummy for each participant-1…
- …instead we assume that this intercept is a normally distributed random variable with a certain variance

What are the consequences?

- For the between subjects manipulation, standard errors may increase significantly!

Results

Output:

```
Random effects:
 Groups    Name        Variance Std.Dev.
 sessionId (Intercept) 1.772    1.331
Number of obs: 15283, groups: sessionId, 493

Fixed effects:
                      Estimate Std. Error z value Pr(>|z|)
(Intercept)           2.023206   0.152286  13.286  < 2e-16 ***
fmessage1            -0.200659   0.215050  -0.933 0.350778
fmessage2            -0.629890   0.204927  -3.074 0.002114 **
fmessage3            -0.708030   0.207900  -3.406 0.000660 ***
fmessage4            -0.231913   0.211854  -1.095 0.273657
percentage           -0.002294   0.001887  -1.216 0.224172
fmessage1:percentage  0.003966   0.002588   1.533 0.125381
fmessage2:percentage  0.008360   0.002422   3.452 0.000556 ***
fmessage3:percentage  0.003009   0.002453   1.227 0.219986
fmessage4:percentage  0.003125   0.002573   1.215 0.224536
```

Even better?

Can we do better?

- Yes; questions are also repeated!
- Again, we could add a dummy variable for each question
- But let's instead add another random intercept

Add random intercept for questionId:

```
model3 <- glmer(decision ~ fmessage*percentage + (1|sessionId) +
(1|questionId), family=binomial, data=fat2)
```

Results

Output:

```
Random effects:
 Groups     Name        Variance Std.Dev.
 sessionId  (Intercept) 4.161    2.040
 questionId (Intercept) 2.437    1.561
Number of obs: 15283, groups: sessionId, 493; questionId, 31

Fixed effects:
                      Estimate Std. Error z value Pr(>|z|)
(Intercept)           2.902810   0.361157   8.038 9.17e-16 ***
fmessage1            -0.299082   0.319965  -0.935 0.349926
fmessage2            -0.951376   0.305720  -3.112 0.001859 **
fmessage3            -1.040422   0.310106  -3.355 0.000793 ***
fmessage4            -0.350746   0.315350  -1.112 0.266034
percentage           -0.001853   0.002304  -0.804 0.421169
fmessage1:percentage  0.003657   0.003150   1.161 0.245569
fmessage2:percentage  0.009889   0.002957   3.344 0.000825 ***
fmessage3:percentage  0.005157   0.002981   1.730 0.083703 .
fmessage4:percentage  0.002917   0.003134   0.931 0.351925
```

INFORMATION AND COMPUTER SCIENCES

Is it better?

Compare (nested) models with ANOVA:

```
anova(model2, model3)
```

Result:

```
       Df   AIC   BIC  logLik deviance  Chisq Chi Df Pr(>Chisq)
model2 11 14018 14102 -6998.0    13996
model3 12 10487 10578 -5231.3    10463 3533.3      1  < 2.2e-16 ***
```

The difference is significant!

Even better?

Can we do better?

- Maybe percentage has a different influence per participant?
- Again, we could add an interaction of percentage*sessionId (lots of dummies!)
- But let's instead add a random slope

Add random slope for percentage and sessionId:

```
model3 <- glmer(decision ~ fmessage*percentage + (1+percentage|sessionId)
+ (1|questionId), family=binomial, data=fat2)
```

Results

Output:

```
Random effects:
 Groups     Name        Variance  Std.Dev. Corr
 sessionId  (Intercept) 4.198e+00 2.048948
            percentage  3.344e-05 0.005783 0.20
 questionId (Intercept) 2.459e+00 1.568018
Number of obs: 15283, groups: sessionId, 493; questionId, 31

Fixed effects:
                      Estimate Std. Error z value Pr(>|z|)
(Intercept)           2.915962   0.362914   8.035 9.37e-16 ***
fmessage1            -0.300180   0.321299  -0.934 0.350165
fmessage2            -0.951371   0.307077  -3.098 0.001947 **
fmessage3            -1.040798   0.311430  -3.342 0.000832 ***
fmessage4            -0.352260   0.316665  -1.112 0.265963
percentage           -0.001280   0.002515  -0.509 0.610853
fmessage1:percentage  0.003833   0.003304   1.160 0.246014
fmessage2:percentage  0.010004   0.003113   3.214 0.001311 **
fmessage3:percentage  0.005100   0.003132   1.628 0.103431
fmessage4:percentage  0.002877   0.003284   0.876 0.380994
```

Is it better?

Compare (nested) models with ANOVA:

```
anova(model3, model4)
```

Result:

```
       Df   AIC   BIC  logLik deviance  Chisq Chi Df Pr(>Chisq)
model3 12 10487 10578 -5231.3    10463
model4 14 10488 10595 -5230.2    10460 2.2451      2     0.3255
```

The difference is **not** significant!

GEE

GEE = General Estimating Equations

Works on normal data and binary/count data

R package: geepack

Function: geeglm

Formula:

```
gee <- geeglm(decision ~ fmessage*percentage, id=sessionId,
family=binomial, corstr="exchangeable", data=fat2)
```

Interpretation

The 15283 data points originate from 493 participants, so errors are correlated within each participant

How do we deal with this?

- We allow correlations in the error covariance matrix
- These errors are allowed to correlate with some equal amount *alpha*

Results

Output:

```
 Coefficients:
                     Estimate  Std.err   Wald Pr(>|W|)
(Intercept)           1.58530  0.13248 143.19  < 2e-16 ***
fmessage1            -0.21784  0.18306   1.42  0.23404
fmessage2            -0.51499  0.16626   9.59  0.00195 **
fmessage3            -0.63060  0.17529  12.94  0.00032 ***
fmessage4            -0.20758  0.17392   1.42  0.23267
percentage           -0.00174  0.00160   1.18  0.27657
fmessage1:percentage  0.00303  0.00209   2.10  0.14755
fmessage2:percentage  0.00664  0.00217   9.34  0.00224 **
fmessage3:percentage  0.00228  0.00203   1.26  0.26228
fmessage4:percentage  0.00239  0.00217   1.21  0.27035

[...]

Estimated Correlation Parameters:
      Estimate Std.err
alpha    0.202  0.0219
Number of clusters:   493   Maximum cluster size: 31
```

Even better?

GLMM can also handle time series data

- Each question is correlated with surrounding questions
- Use "ar1" instead of "exchangeable"
- Specify the order of questions using the "waves" parameter

No examples for this; try it yourself ;-)

Repeated SEM

Can we do this in SEM too?

- Yes! Both ways!

GEE-like SEM

Under VARIABLE:

Specify id variable (cluster = userid)

Under ANALYSIS:

Specify complex model (type = complex)

GEE-like SEM

Advantages:

Simple specification, works just like regular SEM

Disadvantages:

Only two levels; no random slopes or double intercepts

GLMM-like SEM

Under VARIABLE:

Specify within-subjects variables (within = a b c)

Specify between-subjects variables (between = x y z)

Specify id variable (cluster = userid)

Under ANALYSIS:

Specify two-level model (type = twolevel)

Under MODEL:

Specify %within% and %between% effects

GLMM-like SEM

Advantages:

- Can do more than two levels ("threelevel"), and even combine with GEE ("twolevel complex")
- Does intercepts; also random slopes ("twolevel random")
- The random slope can be a dependent variable in another regression (cross-level interactions)

Disadvantages:

- Cannot use categorical indicators
- Can take a long time to estimate (especially "random")

Learn more?

Take a class:

STATS 203

Learn it yourself:

Fitzmaurice, Laird and Ware, "Applied Longitudinal Analysis"

MPlus course videos (the advanced sessions)

Interaction effects

in SEM

Interaction effects

What is the combined effect of x1 and x2 on y?

Possibilities:

- **Additive effect**
- Super-additive effect
- Sub-additive effect
- Cross-over

	x1 = low	x1 = high
x2 = low	0	5
x2 = high	5	10

Interaction effects

What is the combined effect of x1 and x2 on y?

Possibilities:

- Additive effect
- **Super-additive effect**
- Sub-additive effect
- Cross-over

	x1 = low	**x1 = high**
x2 = low	0	5
x2 = high	5	**15**

Interaction effects

What is the combined effect of x1 and x2 on y?

Possibilities:

- Additive effect
- Super-additive effect
- **Sub-additive effect**
- Cross-over

	x1 = low	x1 = high
x2 = low	0	5
x2 = high	5	**5**

Interaction effects

What is the combined effect of x1 and x2 on y?

Possibilities:

- Additive effect
- Super-additive effect
- Sub-additive effect
- **Cross-over**

	x1 = low	**x1 = high**
x2 = low	0	5
x2 = high	5	**0**

Model specification

This is easy in regressions

- Just multiply the dependent variables!
- y ~ x1*x2

More difficult in SEM

- Depends on type of variables:
- manipulation * manipulation
- manipulation * factor
- factor * factor

Model specification

manipulation * manipulation is easy:

- Just create the dummies!
- See SEM slides for an example

manipulation * factor:

- Multiple groups model or predicted random slopes model

factor * factor:

- Predicted random slopes model

Two approaches

"Predicted random slopes model"

- Pro: Works for all types of variables
- Con: Cannot use categorical indicators
- Con: Can take a long time to estimate

"Multiple groups model"

- Pro: Easier to estimate
- Pro: Can sometimes use categorical indicators*
- Con: Does not work for factor * factor interactions

Random slopes

Under ANALYSIS:

Specify random slopes (type = random)

Specify integration (algorithm = integration)

Under MODEL:

Specify the moderated effect as random: s | y on x;

Regress the slope on the moderator: s on m;

Add main effect of moderator: y on m;

Factor * factor

Example: is the effect of perceived control on perceived recommendation quality dependent on understandability?

In regression terms:

quality ~ control*underst

In SEM:

s | quality ON control;

s ON underst;

quality ON underst;

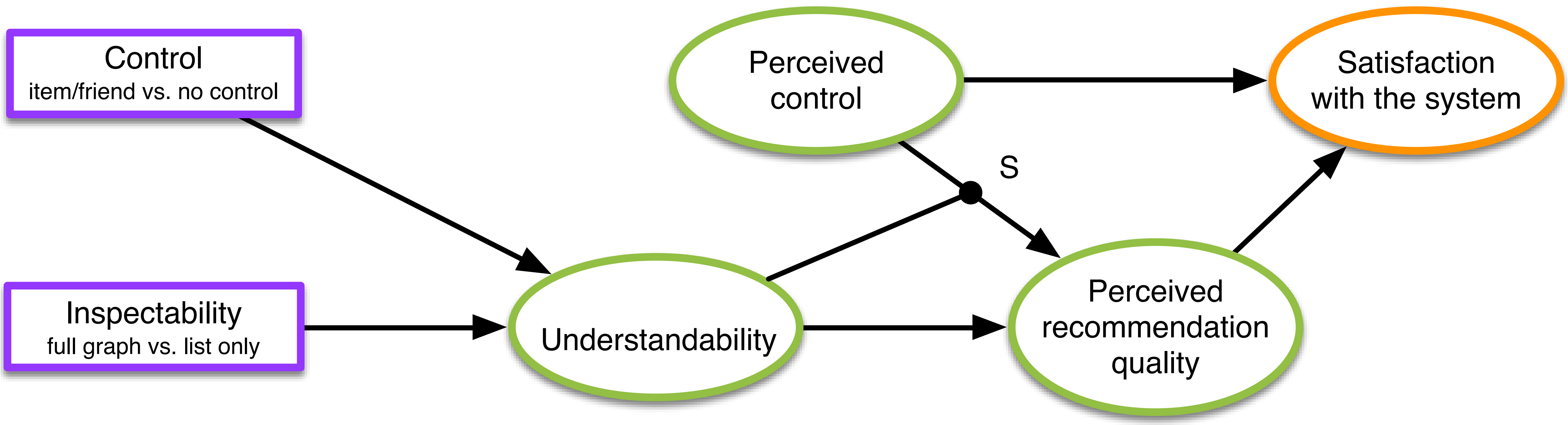
Factor * factor

```
ANALYSIS:
 type = random;
 algorithm = integration;

MODEL:
 satisf BY s1* s2-s7;
 quality BY q1* q2-q6;
 control BY c1* c2-c4;
 underst BY u2* u4-u5;
 satisf-underst@1;

 satisf ON quality control;
 s | quality ON control;
 s ON underst;
 quality ON underst;
 underst ON citem cfriend cgraph;
```

Factor * factor



Factor * condition

Example: is the effect of perceived control on perceived recommendation quality dependent on the control condition?

In SEM:

s | quality ON control;

s ON citem cfriend;

quality ON citem cfriend;

Factor * condition

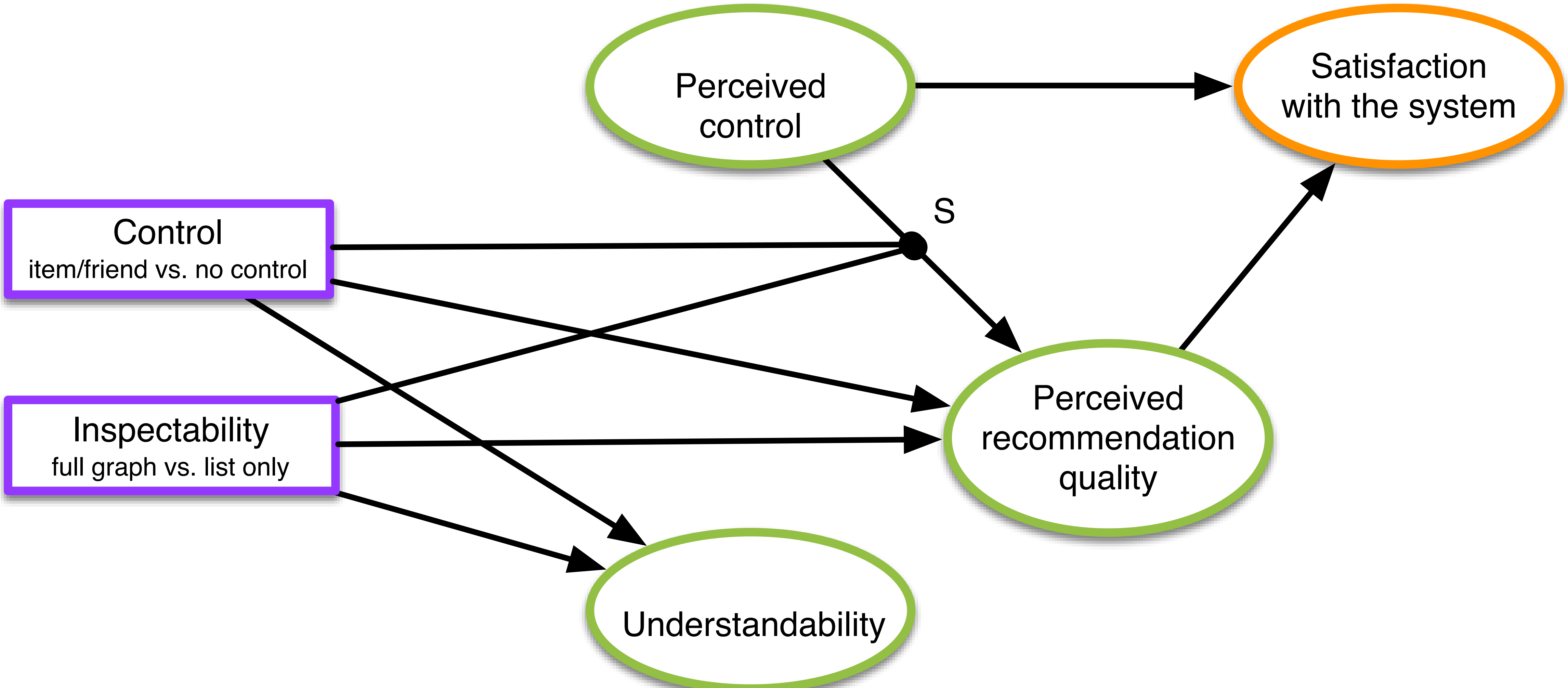
```
ANALYSIS:
 type = random;
 algorithm = integration;

MODEL:
 satisf BY s1* s2-s7;
 quality BY q1* q2-q6;
 control BY c1* c2-c4;
 underst BY u2* u4-u5;
 satisf-underst@1;

 satisf ON quality control;
 s | quality ON control;
 s ON citem cfriend;
 quality ON citem cfriend;
 underst ON citem cfriend cgraph;
```

BREN:ICS INFORMATION AND COMPUTER SCIENCES

Factor * condition



Multiple groups

Under VARIABLE:

Specify the moderating manipulation as a "grouping" variable: grouping = cctrl(0=none 1=item 2=friend)

Add a MODEL section for all groups except the baseline

Model item:

Model friend:

Add corresponding labels to each MODEL to restrict the moderation

Factor * condition

```
MODEL:
 satisf BY s1* s2-s7;
 quality BY q1* q2-q6;
 control BY c1* c2-c4;
 underst BY u2* u4-u5;
 satisf-underst@1;

 satisf ON quality control (1-2);
 quality ON control (p1);
 control ON underst (4);
 underst ON cgraph (5);

 [satisf] (6);
 [quality] (7);
 [control] (8);
 [underst];
```

```
MODEL item:
 satisf ON quality control (1-2);
 quality ON control (p2);
 control ON underst (4);
 underst ON cgraph (5);

 [satisf] (6);
 [quality] (7);
 [control] (8);
 [underst];

MODEL friend:
 satisf ON quality control (1-2);
 quality ON control (p3);
 control ON underst (4);
 underst ON cgraph (5);

 [satisf] (6);
 [quality] (7);
 [control] (8);
 [underst];
```

Learn more?

Learn it yourself:

- MPlus course videos (the advanced sessions)

Cluster Analysis

using Latent Categorical Analysis and
Mixture Factor Analysis

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: Mixture Factor 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~~

LCA

Under VARIABLE:

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

Under ANALYSIS:

Specify mixture model: type = mixture

Optionally, specify iterations etc

MFA

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 :-)

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: tech4)
- **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.

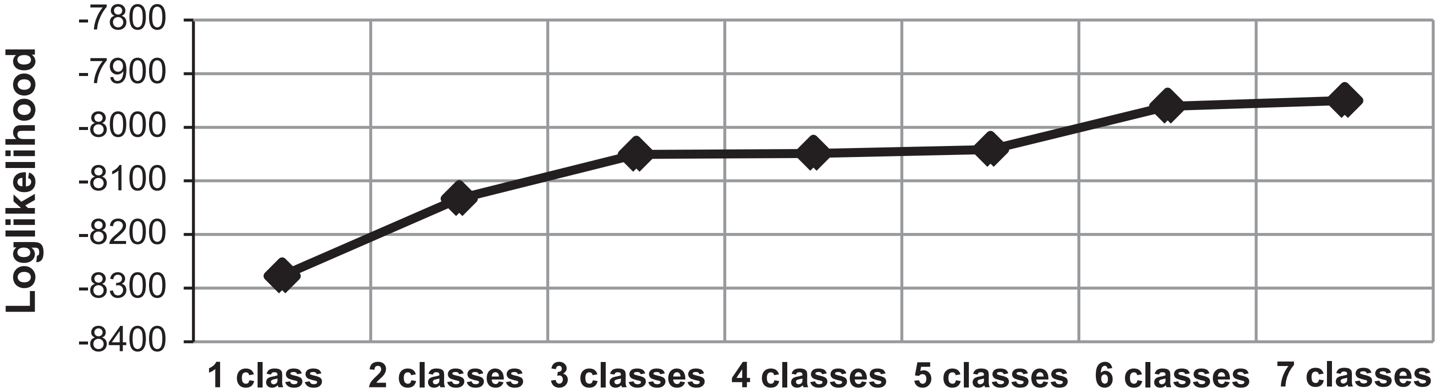
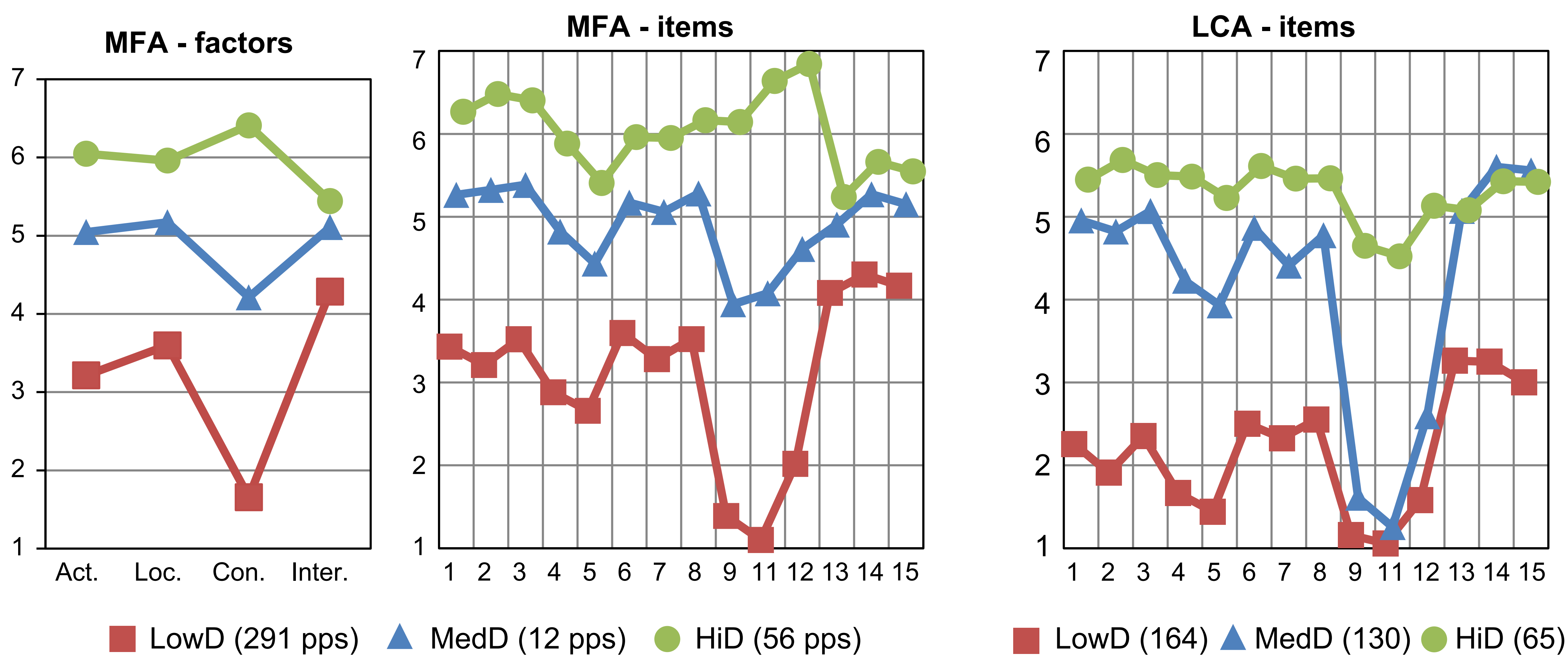


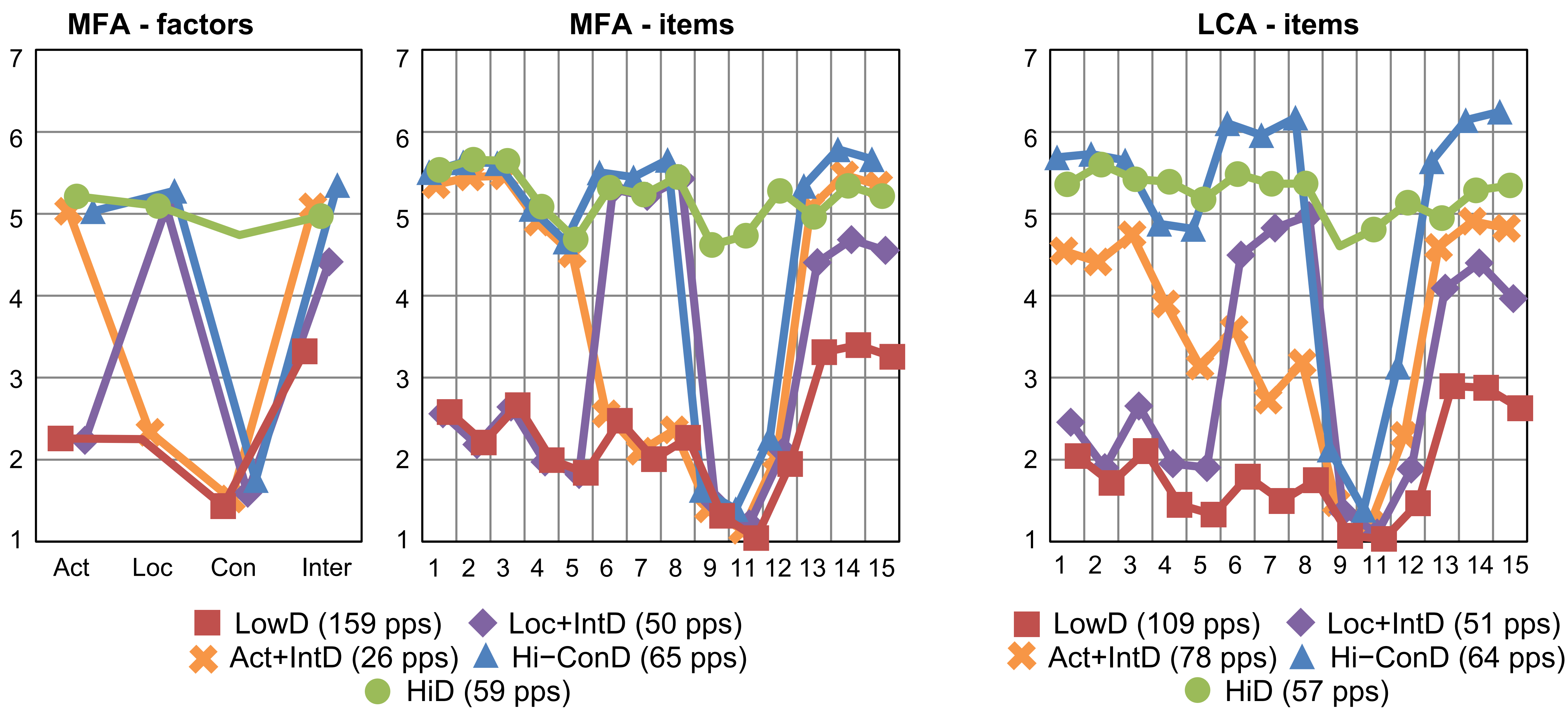
Fig. 8. Change in loglikelihood between subsequent MFA models.

Results



MFA - factors
MFA - items
LCA - items
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 (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)

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

T H A N K S !

George Bernard Shaw