

Item Response Theory (IRT) & Structural Equation Modelling (SEM)

Aleksandar Baucal
abaucal@f.bg.ac.rs

Univ. of Belgrade, Dept. of Psychology



FOCUS
FOOD CONSUMER SCIENCE IN THE BALKANS



Two practical objectives in empirical studies

- 1. To create good scale for measurement of some variable

“New” tool: IRT

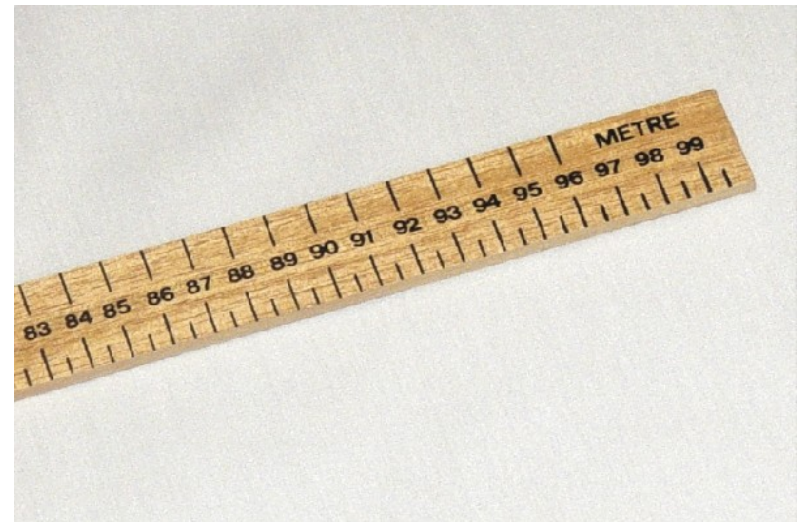
- 2. To sort out how different variables are mutually related

“New” tool: SEM

Item Response Theory (IRT)

Key purpose

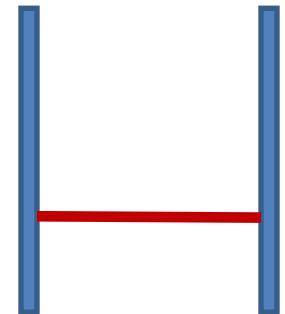
- To enable researchers to develop good measurement scale for some variable, something like “meter stick” for this variable
 - Interval scale
 - Independent from object of measurement
 - Scales equivalence



IRT promise 1

- Problem 1: Classical psychometric model has troubles when it deals with developmental items (items with different level of difficulties)

		Item 2 ($p = 0.4$) Difficult, middle level		
		0	1	Tot
Item1 ($p = 0.8$) Easy, low level	0	20	0	20
	1	40	40	80
	Tot	60	40	100



- Bayesian statistics which is in the basis of the IRT can deal with such set of items since it is based on conjoint probabilities, not on linear correlation

IRT promise 2

- In the classical psychometric approach estimation of subjects' abilities and item difficulties are mutually dependant (it is not easy to generalize estimations)
- In Bayesian model estimation of subjects' abilities and item difficulties are independent and can be generalized
- For example: if the sample is divided in two groups (low achievers vs. high achievers)
 - classical psychometric approach will give us different item parameters (item difficulty and item-total score correlation)
 - IRT gives the same item parameters

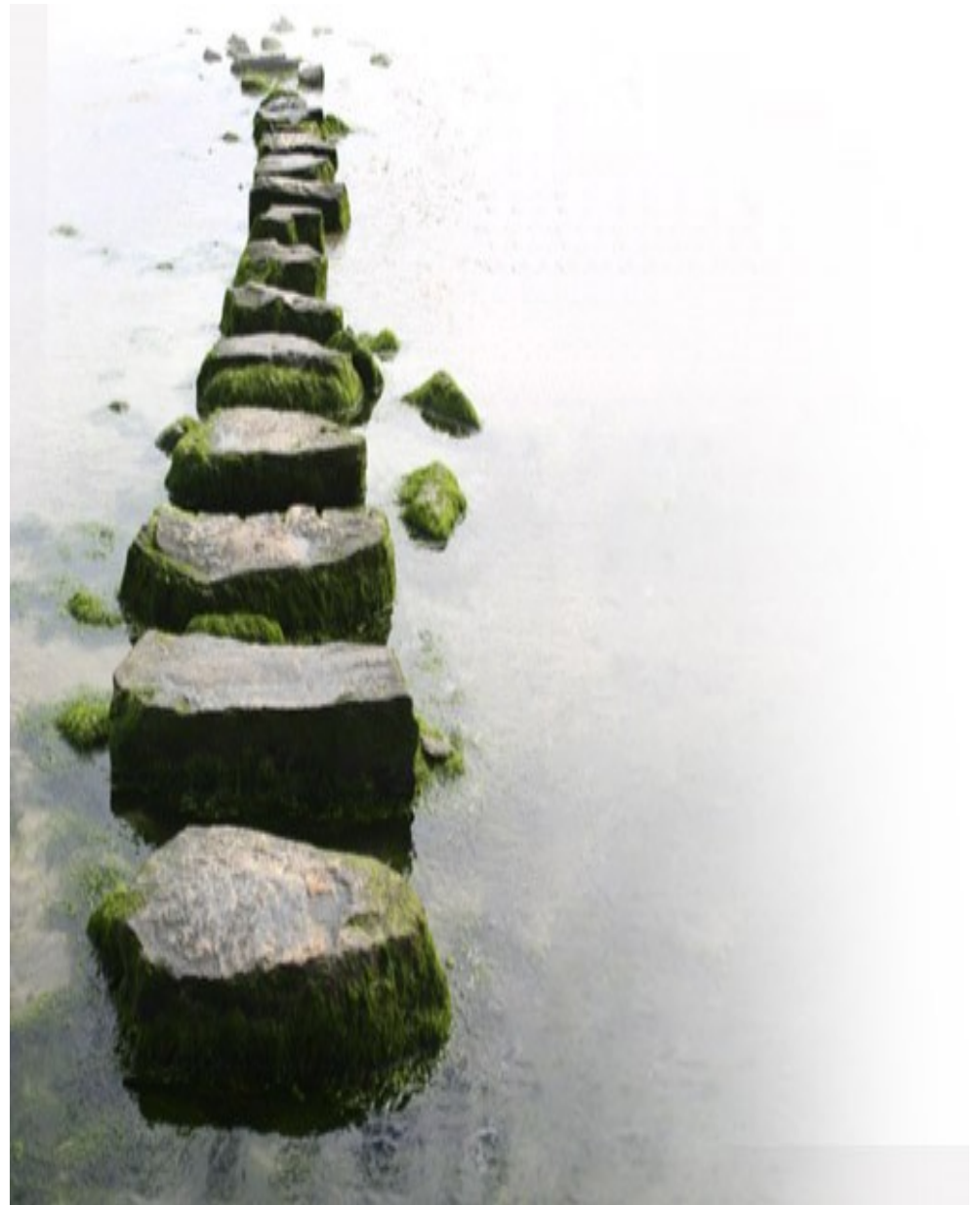
low ability group			high ability group		
Subj.	Q8	Total	Subj.	Q8	Total
1	0	8	16	1	33
2	0	12	17	0	28
3	0	6	18	1	29
4	0	12	19	1	30
5	0	8	20	1	29
6	0	8	21	0	28
7	0	8	22	1	33
8	0	11	23	1	32
9	1	13	24	1	32
10	0	4	25	1	33
11	1	14	26	0	34
12	1	13	27	1	35
13	0	10	28	1	34
14	0	9	29	1	38
15	0	8	30	1	37

Low ability group:
 $p=(3/15)=0.2$
 $r= 0.68$

High ability group:
 $p=(12/15)=0.8$
 $r= 0.39$

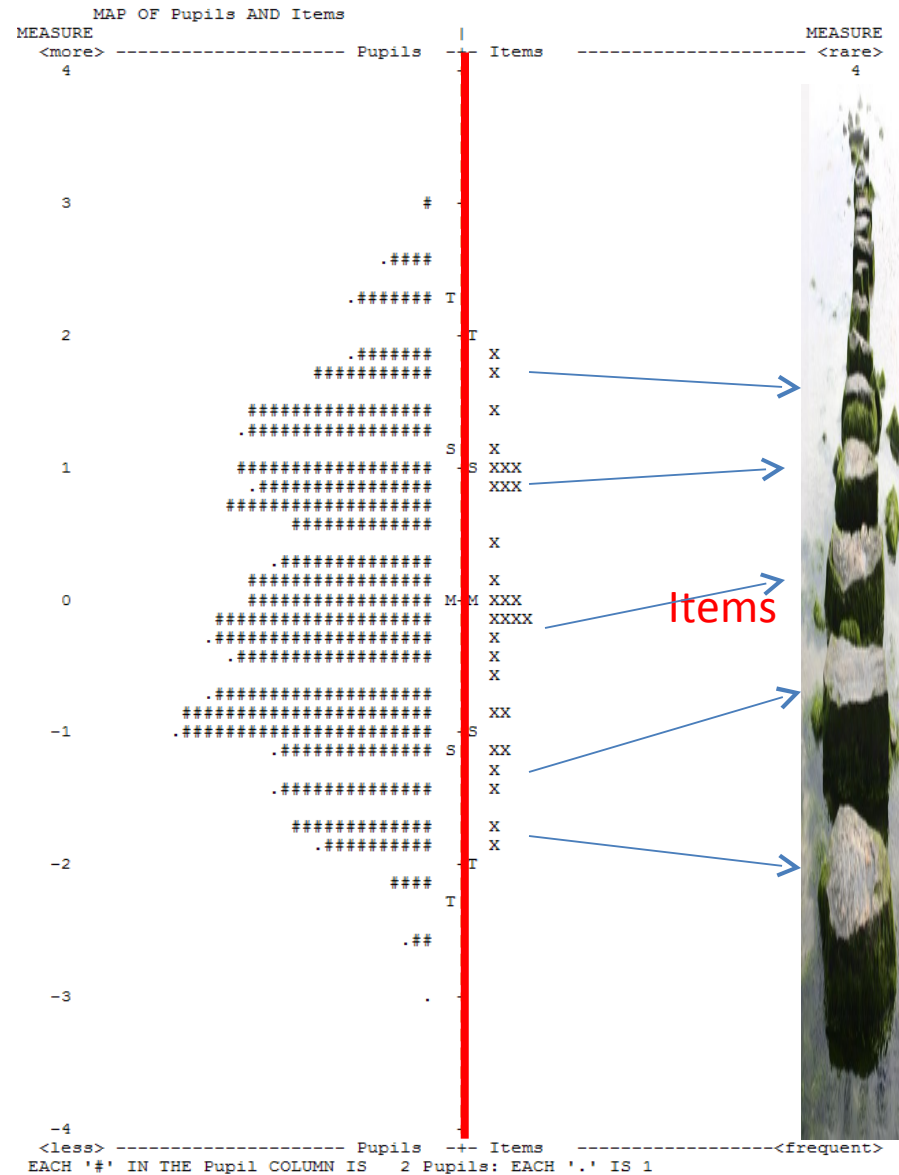
IRT promise 3

- IRT test
unidimensionality



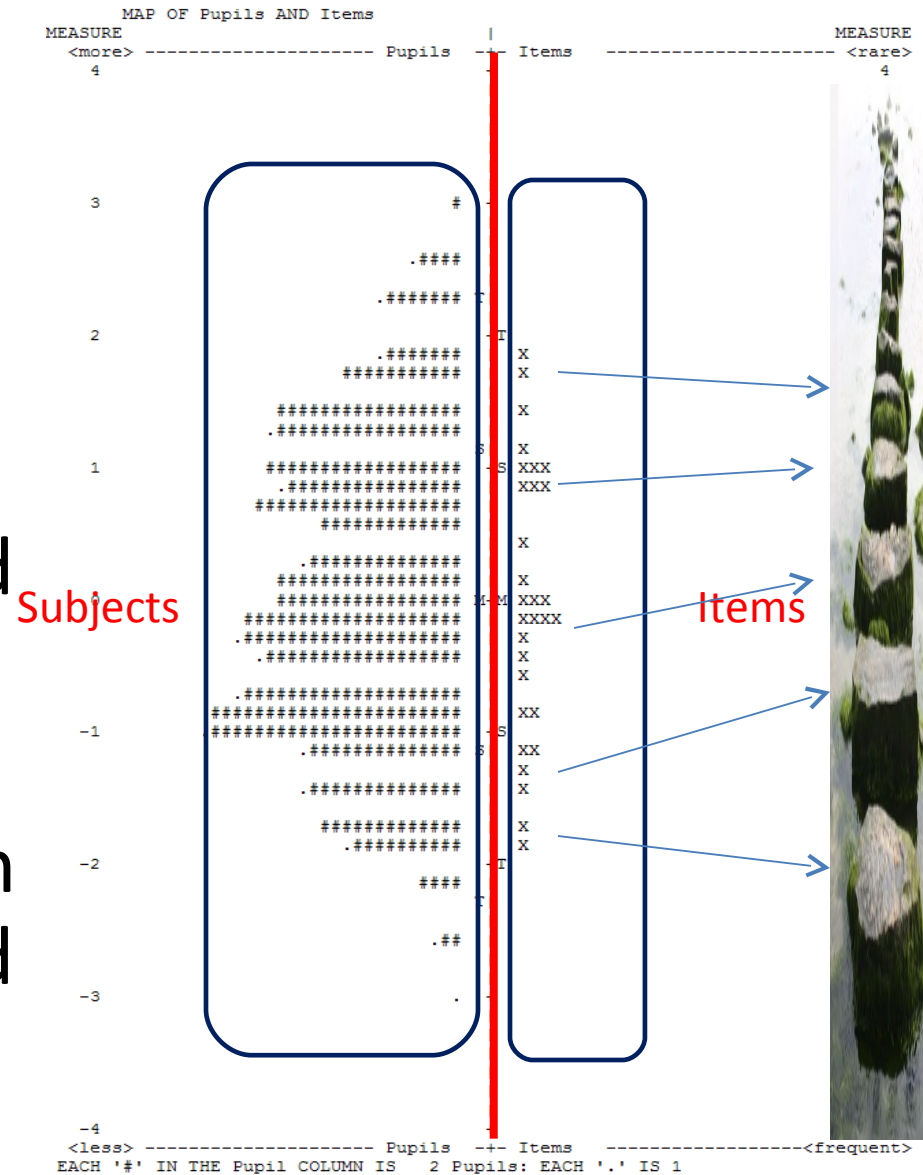
IRT promise 3

- IRT test unidimensionality



IRT promise 3

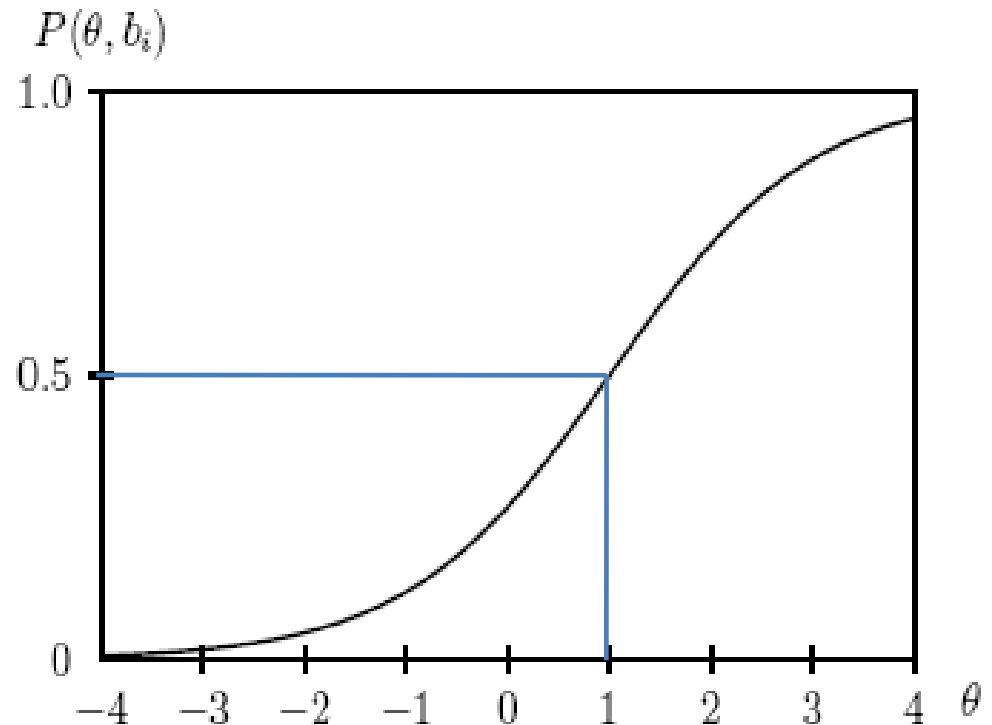
- IRT test unidimensionality
- IRT estimates subjects' abilities and item difficulties at the same scale
- So, it is easy to match subjects' abilities and item difficulties



Item Characteristic Curve (ICC): Core idea

Key assumption:

Relationship between ability of person and probability that this person will solve some item can be described by S curve

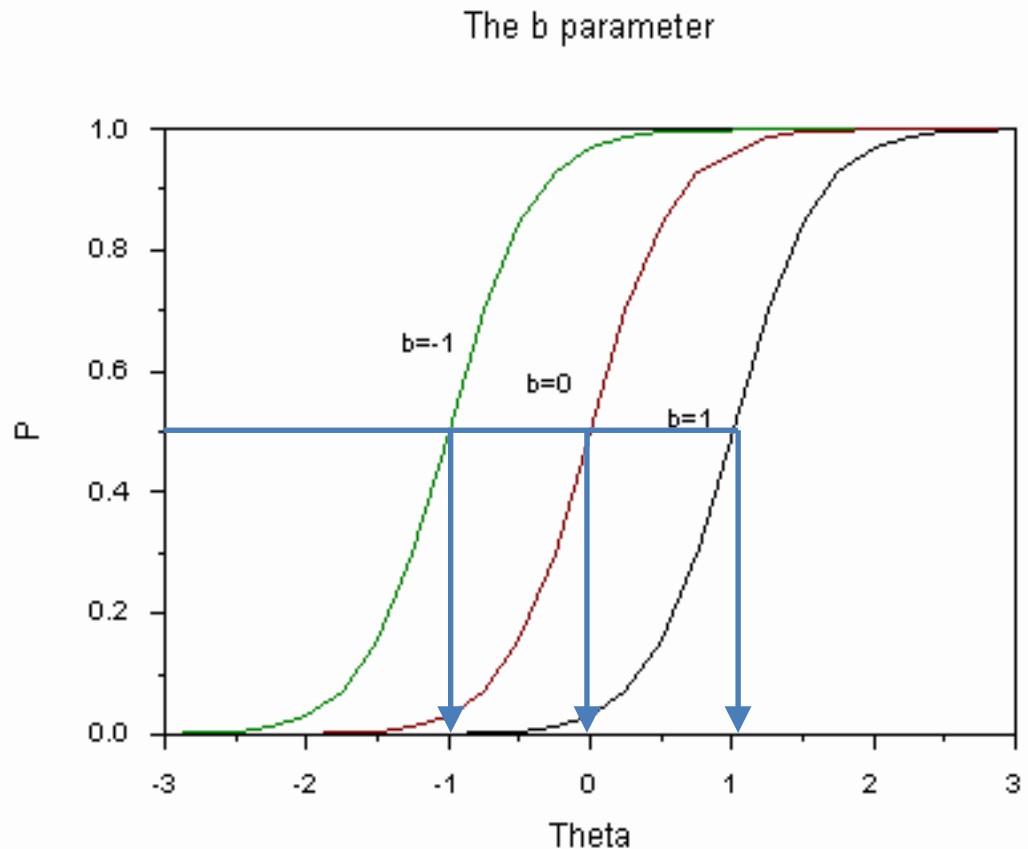


Three parameters of ICC

- **Difficulty of item (b)** - what should be ability level of person to have 50% of chance to solve this item, i.e. what ability level person needs to have in order to be able to solve this item
- **Item discrimination (a)** - steepness of curve, i.e. how probability that one will get item right increase with increasing of person ability
- **Guessing (c)** – what is probability that person will get item right when person ability is far below of item difficulty

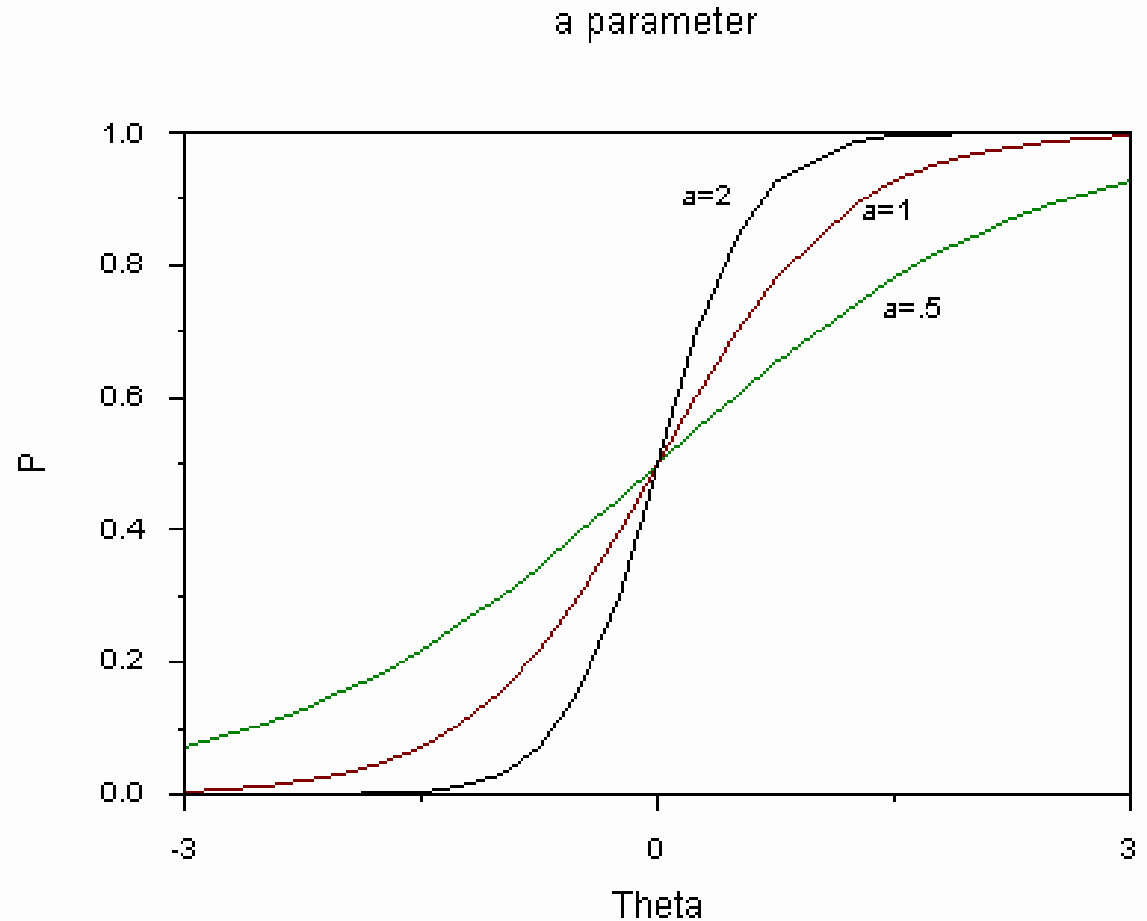
Item difficulty (b parameter)

- Items have different difficulty = one needs to have different ability to have 50% chance to solve different items



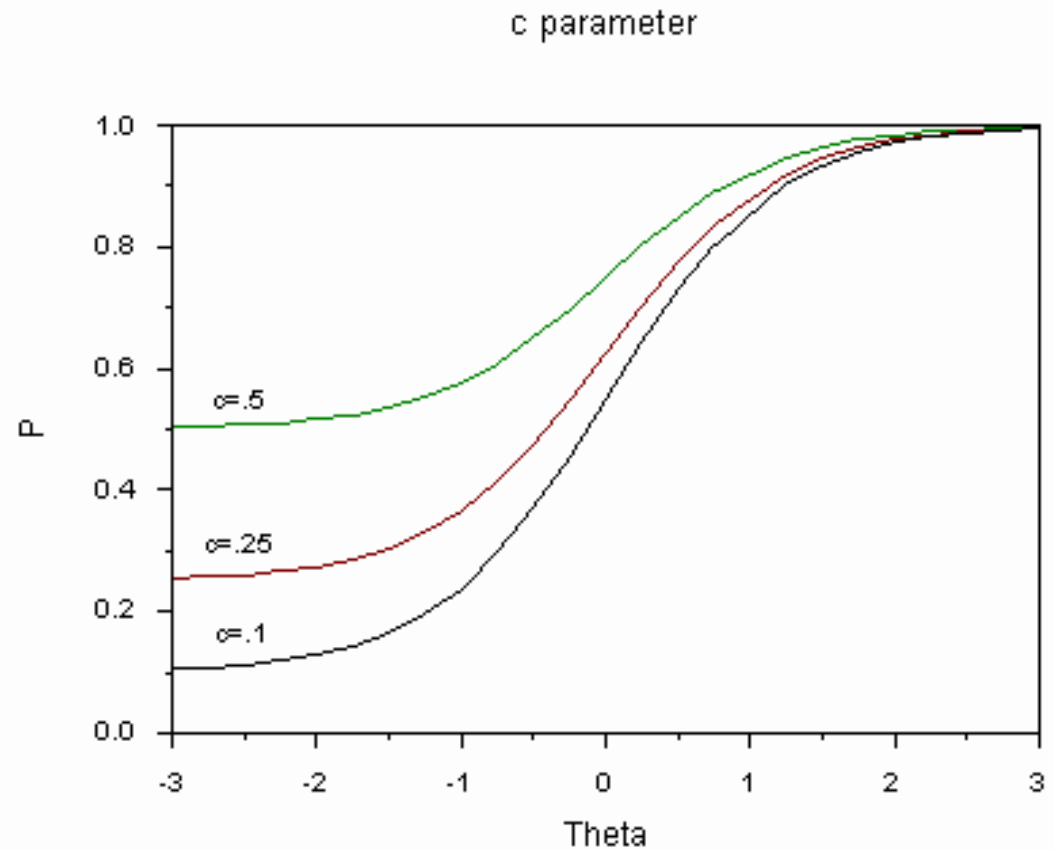
Item discrimination (a parameter)

- Steepness of ICC
- Items with different steepness have different power to discriminate persons with similar ability

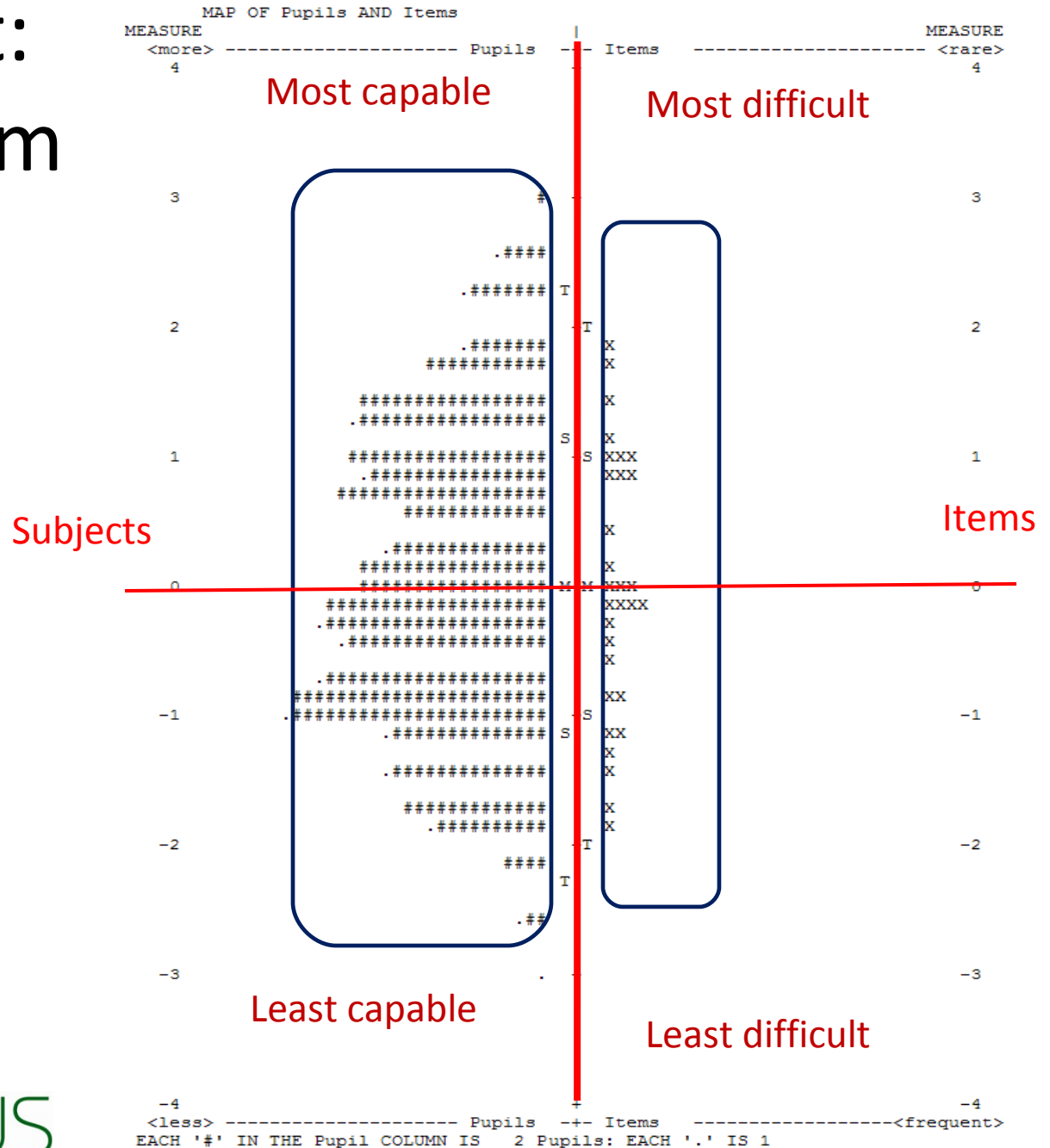


Guessing parameter (c parameter)

- It indicates what is probability that person who do not have needed level of ability can get it right
- Guessing parameter is very useful with “closed items”



Key output: Student/item map



IRT Model test

- Test of the assumption of unidimensionality (all items measure the same construct, items are ordered as developmental milestones)
 - Reliability of item difficulties $> .85-.90$
 - Whether item difficulties would be reproduced when other subject sample would be studied
 - Reliability of subjects' abilities $> .80$
 - Whether person abilities would be reproduced when other set of items would be studied
 - Global infit & outfit measures ($0.7 < MNSQ < 1.3$)
 - Whether items and persons fit the unidimensional model

IRT Model test

SUMMARY OF 733 MEASURED Pupils

	RAW		MEASURE	MODEL ERROR	INFIT		OUTFIT	
	SCORE	COUNT			MNSQ	ZSTD	MNSQ	ZSTD
MEAN	14.9	30.0	.00	.44	.98	-.1	1.11	.2
S.D.	6.1	.0	1.12	.05	.17	.9	.42	1.1
MAX.	28.0	30.0	3.05	.75	1.56	2.9	3.20	3.5
MIN.	2.0	30.0	-3.03	.40	.60	-2.9	.41	-2.7
REAL RMSE	.46	ADJ.SD	1.03	SEPARATION	2.23	Pupil	RELIABILITY	.83
MODEL RMSE	.45	ADJ.SD	1.03	SEPARATION	2.31	Pupil	RELIABILITY	.84
S.E. OF Pupil MEAN = .04								

SUMMARY OF 30 MEASURED Items

	RAW		MEASURE	MODEL ERROR	INFIT		OUTFIT	
	SCORE	COUNT			MNSQ	ZSTD	MNSQ	ZSTD
MEAN	363.8	733.0	.01	.09	.99	-1.4	1.11	-.4
S.D.	132.5	.0	1.00	.01	.24	4.0	.51	4.1
MAX.	600.0	733.0	1.83	.10	2.02	9.9	3.20	9.9
MIN.	136.0	733.0	-1.86	.08	.73	-8.7	.69	-7.2
REAL RMSE	.09	ADJ.SD	1.00	SEPARATION	10.73	Item	RELIABILITY	.99
MODEL RMSE	.09	ADJ.SD	1.00	SEPARATION	11.15	Item	RELIABILITY	.99
S.E. OF Item MEAN = .19								

Test of each item

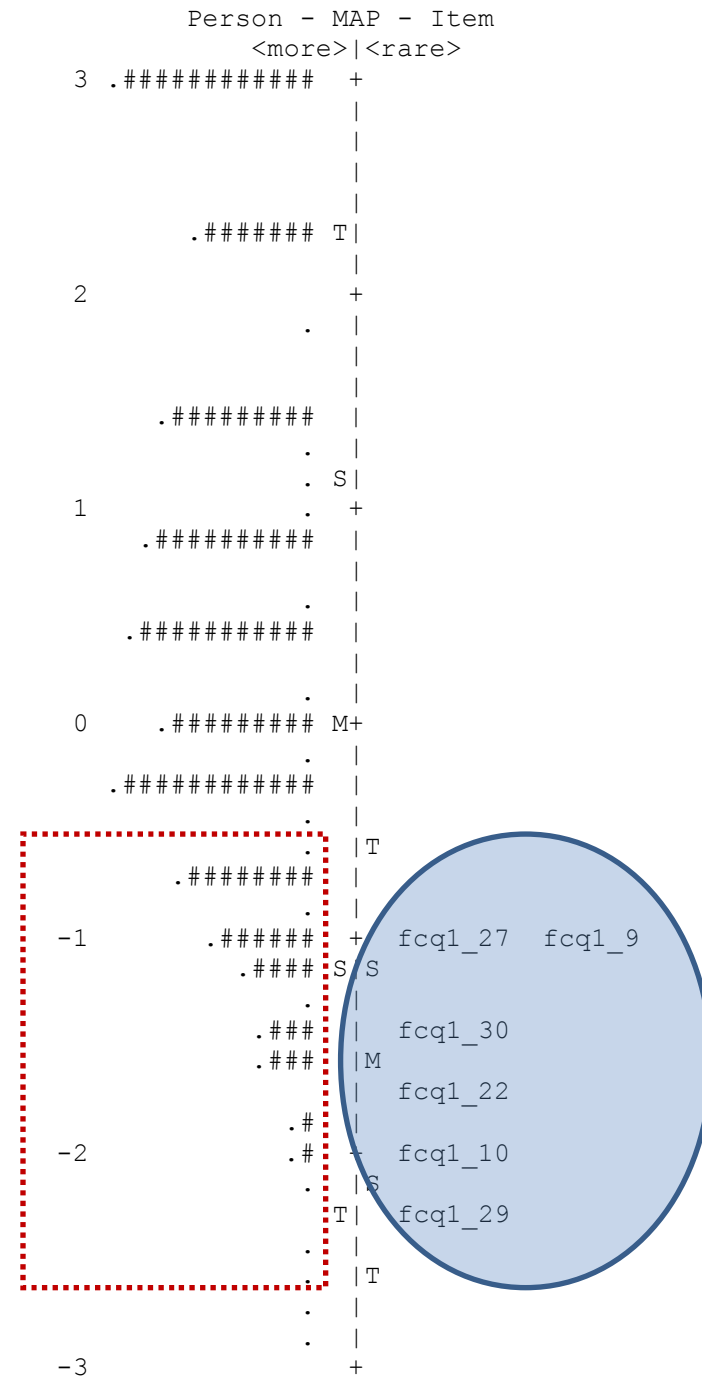
ENTRY NUMBER	RAW SCORE	COUNT	MEASURE	MODEL S.E.	INFIT MNSQ ZSTD	OUTFIT MNSQ ZSTD	PTMEA CORR.	EXACT OBS%	MATCH EXP%	Item		
11	225	733	1.02	.09	2.02	9.9	3.20	9.9	A-.49	48.8	74.8	v1q11
20	600	733	-1.86	.10	1.27	4.3	2.07	7.0	B .05	81.2	82.4	v1q20
10	143	733	1.76	.10	1.37	5.9	1.95	7.0	C .04	78.2	81.7	v1q10
25	521	733	-1.14	.09	1.34	7.6	1.75	7.9	D .10	66.6	74.7	v1q25
15	580	733	-1.66	.10	1.14	2.6	1.38	3.2	E .23	78.4	80.1	v1q15
7	177	733	1.43	.10	.99	-.2	1.18	1.9	F .39	80.1	78.3	v1q7
28	522	733	-1.15	.09	.99	-.3	1.11	1.4	G .41	77.4	74.8	v1q28
22	136	733	1.83	.10	.94	-1.0	1.10	.9	H .40	84.6	82.5	v1q22
6	219	733	1.07	.09	.97	-.7	1.08	1.1	I .44	77.2	75.1	v1q6
24	553	733	-1.41	.09	.93	-1.5	1.06	.7	J .42	81.7	77.4	v1q24
17	341	733	.17	.08	1.01	.4	.99	-.1	K .46	70.1	70.5	v1q17
5	448	733	-.58	.08	.98	-.6	1.01	.2	L .45	73.0	70.9	v1q5
30	493	733	-.92	.09	.94	-1.7	1.01	.2	M .46	77.6	72.8	v1q30
21	484	733	-.85	.09	.94	-1.6	.97	-.5	N .47	76.4	72.3	v1q21
13	408	733	-.30	.08	.96	-1.1	.93	-1.3	O .49	70.5	70.1	v1q13
3	250	733	.82	.09	.94	-1.5	.96	-.6	o .49	77.2	73.4	v1q3
29	232	733	.96	.09	.92	-1.9	.90	-1.5	n .50	78.2	74.4	v1q29
14	254	733	.79	.09	.92	-2.1	.92	-1.4	m .51	77.2	73.2	v1q14
2	531	733	-1.22	.09	.88	-3.1	.85	-2.0	l .50	78.7	75.5	v1q2
18	362	733	.02	.08	.87	-3.9	.82	-3.9	k .57	74.9	70.2	v1q18
26	360	733	.03	.08	.87	-4.1	.85	-3.4	j .56	76.0	70.2	v1q26
23	202	733	1.21	.09	.85	-3.6	.86	-1.9	i .54	81.2	76.3	v1q23
19	433	733	-.48	.08	.86	-4.3	.82	-3.6	h .56	77.1	70.5	v1q19
4	299	733	.46	.08	.86	-4.2	.82	-3.7	g .57	76.5	71.5	v1q4
12	389	733	-.17	.08	.85	-4.6	.82	-4.0	f .58	76.9	70.0	v1q12
27	357	733	.05	.08	.85	-4.7	.81	-4.3	e .58	75.9	70.2	v1q27
1	242	733	.89	.09	.85	-4.0	.83	-2.8	d .56	80.1	73.8	v1q1
8	390	733	-.18	.08	.81	-6.1	.77	-5.1	c .61	79.5	70.0	v1q8
16	382	733	-.12	.08	.79	-6.7	.74	-5.9	b .63	79.5	70.0	v1q16
9	380	733	-.11	.08	.73	-8.7	.69	-7.2	a .67	82.8	70.1	v1q9
MEAN	363.8	733.0	.01	.09	.99	-1.4	1.11	-.4		76.5	73.9	
S.D.	132.5	.0	1.00	.01	.24	4.0	.51	4.1		6.3	3.9	

IRT and Likert scale data

- Previous examples are taken from the field of cognitive assessment where in the most cases person performance is described with 0 (failed) and 1 (succeed)
- In the case of attitudes researchers usually ask participants to assess their agreement with some statements on Likert scale (1-5)
- In this case IRT analysis assess place of each score of the Likert scale on the measurement scale

Example from FCQ questions (health subscale)

- Scale is rather easy for representative sample
- Can differentiate only persons on the lower part of the scale, not on the upper part



Example from FCQ questions (health subscale)

Item STATISTICS: MEASURE ORDER

ENTRY NUMBER	TOTAL SCORE	COUNT	MEASURE	MODEL S.E.	INFIT MNSQ	INFIT ZSTD	OUTFIT MNSQ	OUTFIT ZSTD	PT-MEASURE CORR.	EXACT MATCH EXP.	EXACT MATCH OBS%	EXACT MATCH EXP%	Item	G
9	11790	3069	-0.95	.03	1.17	5.7	1.21	7.2	.66	.71	50.3	51.2	fcq1_9	0
27	11874	3067	-1.00	.03	1.08	2.9	1.12	4.1	.68	.70	52.8	50.6	fcq1_27	0
30	12881	3069	-1.50	.03	.92	-2.7	.94	-2.0	.65	.63	58.2	53.0	fcq1_30	0
22	13239	3074	-1.76	.03	.87	-4.1	.87	-4.1	.64	.61	62.3	56.8	fcq1_22	0
10	13328	3071	-2.05	.03	1.11	3.3	1.12	3.6	.58	.60	60.5	58.6	fcq1_10	0
29	13655	3073	-2.26	.03	.84	-5.2	.79	-6.5	.62	.57	66.3	60.3	fcq1_29	0
MEAN	12794.5	3070.5	-1.59	.03	1.00	.0	1.01	.4			58.4	55.1		
S.D.	717.2	2.6	.49	.00	.13	4.1	.15	4.9			5.4	3.7		

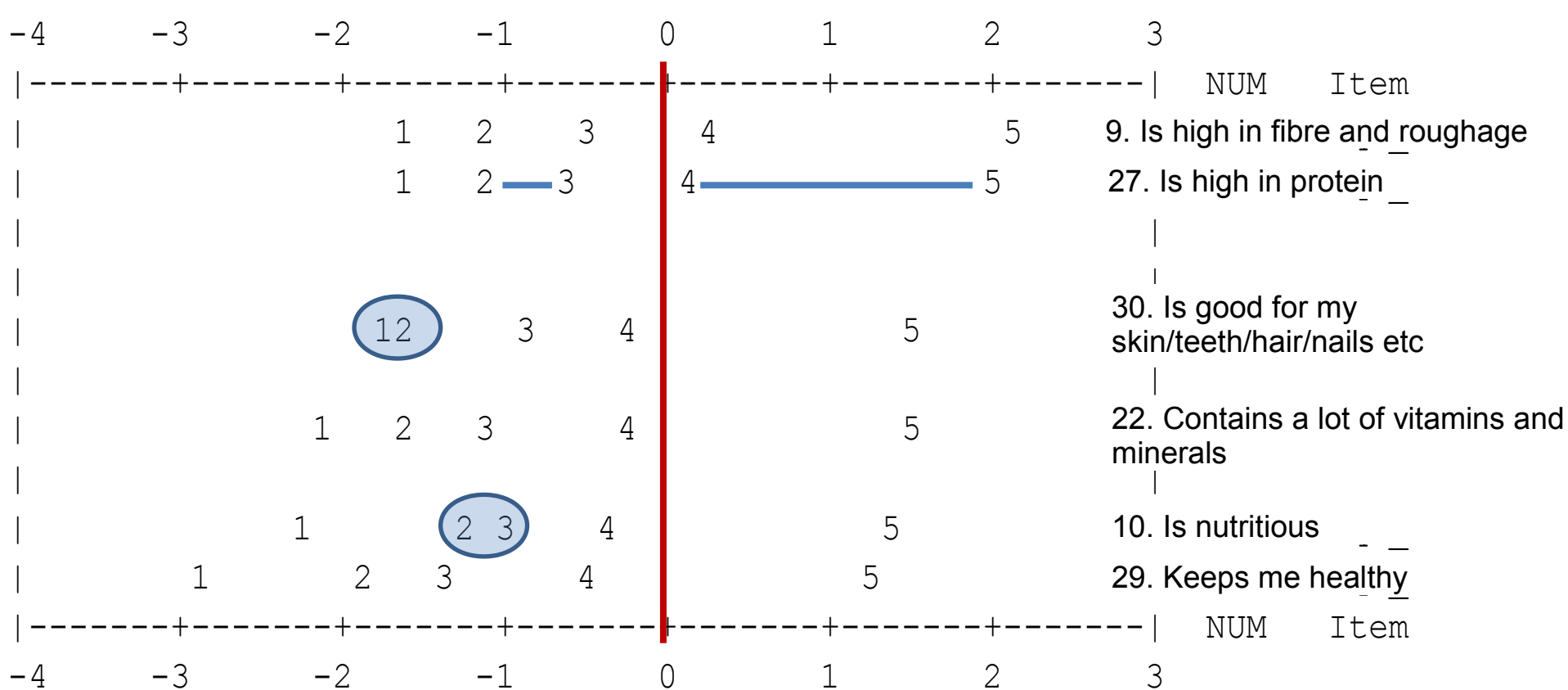
9. Is high in fibre and roughage
27. Is high in protein

30. Is good for my skin/teeth/hair/nails etc
10. Is nutritious

22. Contains a lot of vitamins and minerals
29. Keeps me healthy

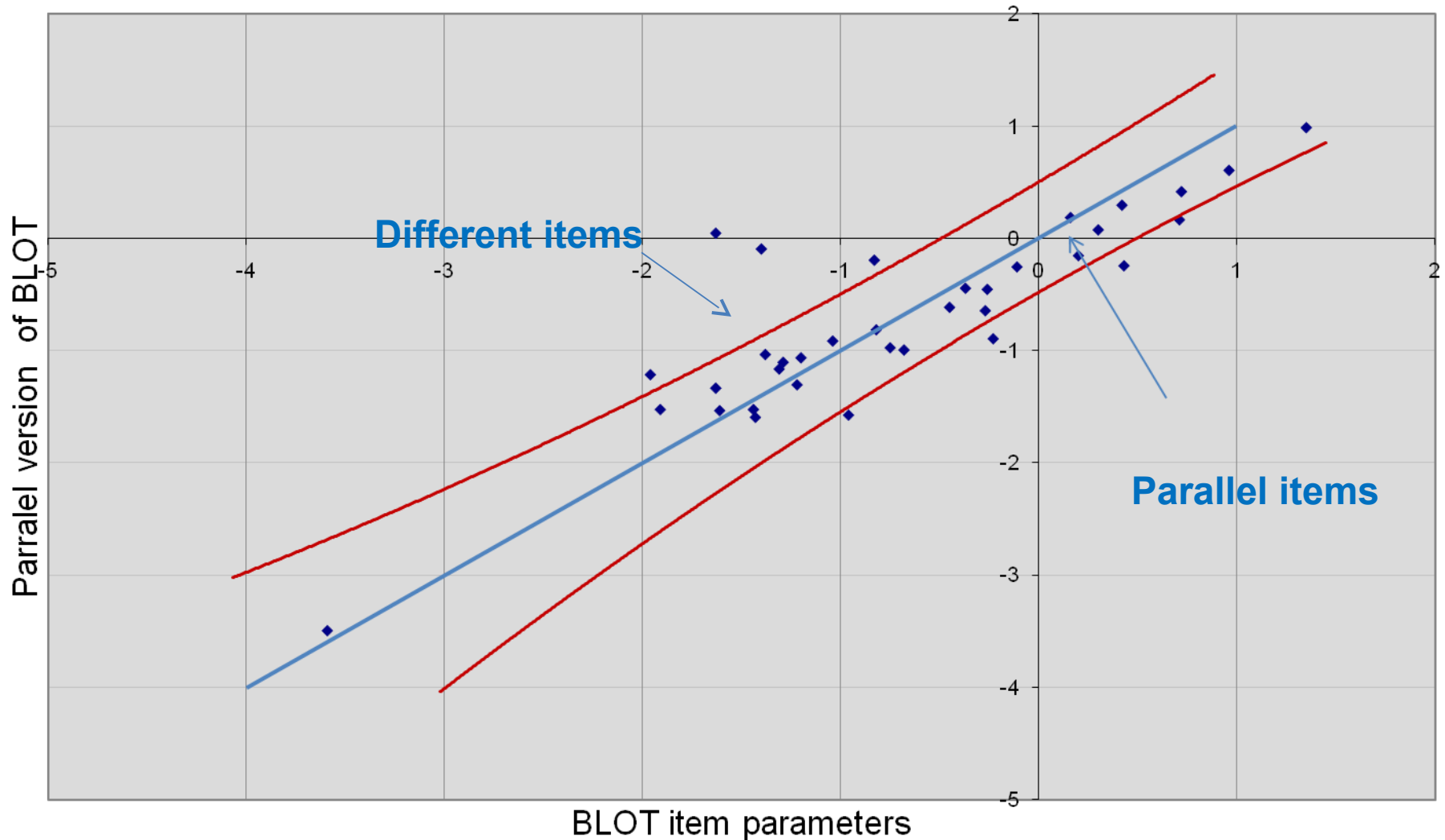
Example from FCQ questions (health subscale)

OBSERVED AVERAGE MEASURES FOR Person (scored)



Very useful application: equating different scales of the same construct

Comparison of item parameters for two versions of BLOT test



Another useful application: to check whether items are “culture fair”

- “Culture fair” means whether participants from different social and cultural groups who have same ability have same probability to get it right (or to give same answer on the Likert scale)
- If some items using for measurement of some construct are not culture fair it can lead toward biased measurement
- Typical examples: success of boys and girls on some cognitive items with typical male or female content

Example from FCQ questions (health subscale)

DIF class specification is: DIF=\$S10W1

Person	DIF	DIF	Person	DIF	DIF	DIF	JOINT	Welch	MantelHanzl	Item				
CLASS	MEASURE	S.E.	CLASS	MEASURE	S.E.	CONTRAST	S.E.	t	d.f.	Prob.	Prob.	Size	Number	Name
1	-.63	.06	2	-.98	.06	.36	.09	3.98	893	.0001	.0004	.30	9	fcq1_9
1	-.63	.06	3	-1.07	.06	.44	.09	4.96	925	.0000	.0000	.33	9	fcq1_9
1	-.63	.06	4	-1.00	.07	.37	.09	4.03	879	.0001	.0000	.30	9	fcq1_9
1	-.63	.06	5	-.98	.07	.35	.09	3.70	840	.0002	.0005	.32	9	fcq1_9
1	-.63	.06	6	-1.07	.06	.44	.09	5.00	906	.0000	.0000	.41	9	fcq1_9

Some IRT software

- Winsteps: www.winsteps.com
- Conquest:
<http://www.assess.com/xcart/product.php?product>
- Bilog, Multilog, Parscale
www.ssicentral.com/irt/index.html

Group work

- Divide yourself in smaller groups
- Select one variable from FOCUS questionnaire and its indicators for which you believe that IRT analysis should be done
- Formulate prediction what indicator could be easier or more difficult to endorse

Structural Equation Modelling (SEM)

Have you ever heard?

- It is not possible to draw causal conclusions from correlations?
- What is the basic reason for this?

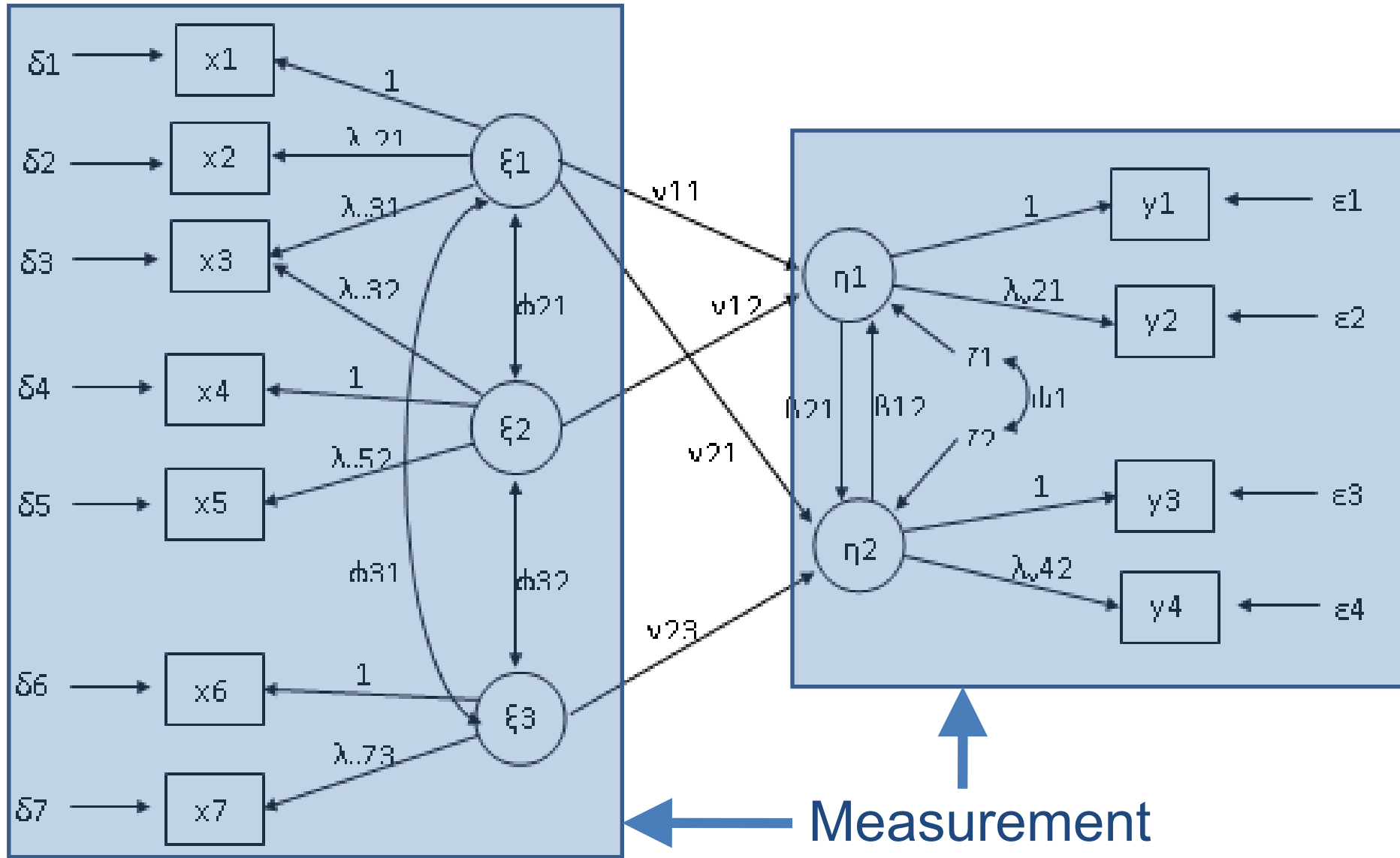
And then at the end of 60s...

- Somebody comes and make a claim that there is a way to formulate causal statements based on correlation data
- This is why the SEM was developed

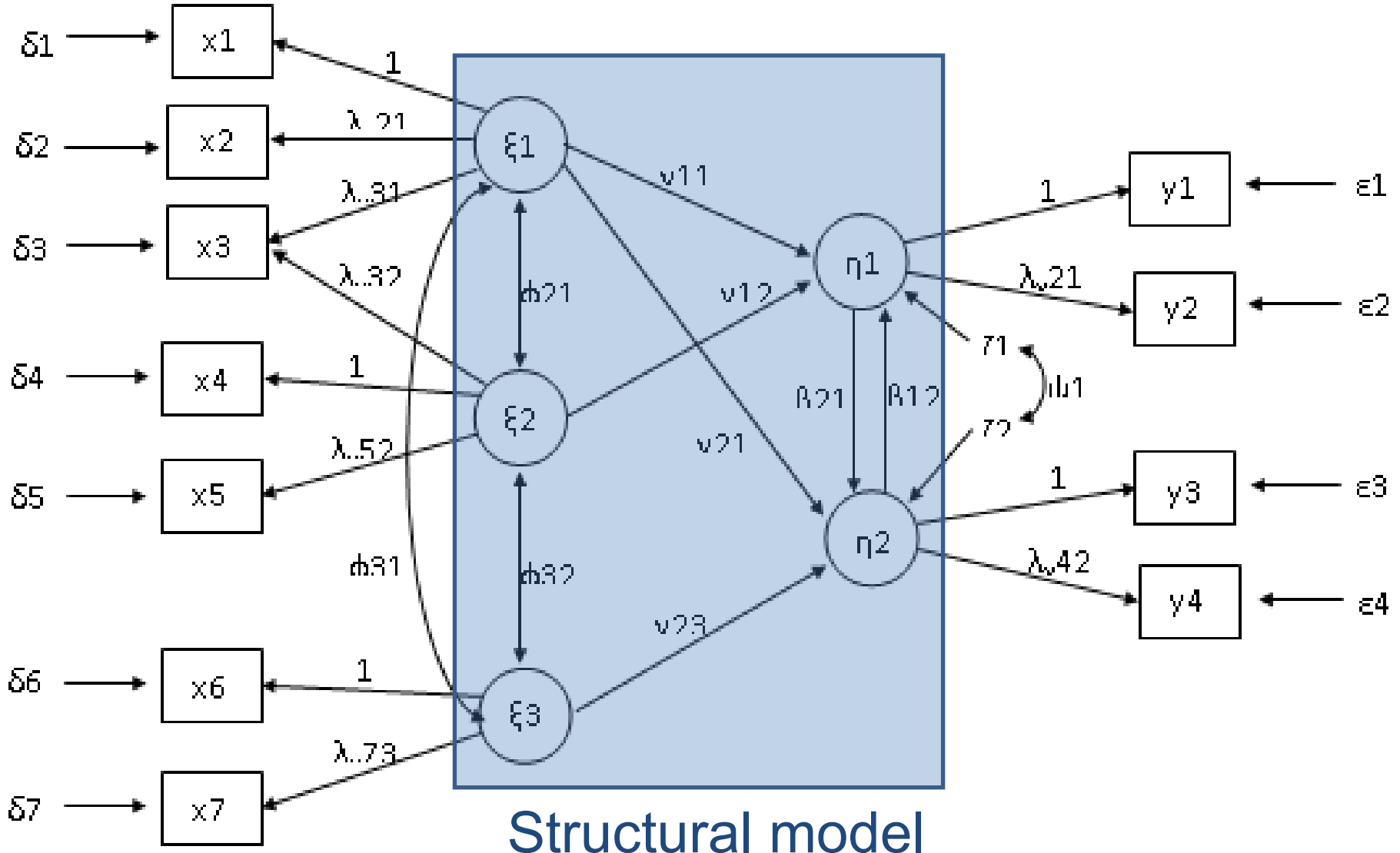
Basic purpose of SEM analysis

- To test validity of a model consisting of causal and covariance relations between latent and manifest variables.
- However, if the model is supported by the SEM analysis it does not mean that it is proved as only one model. Some other models can fit data as well.

Full SEM model

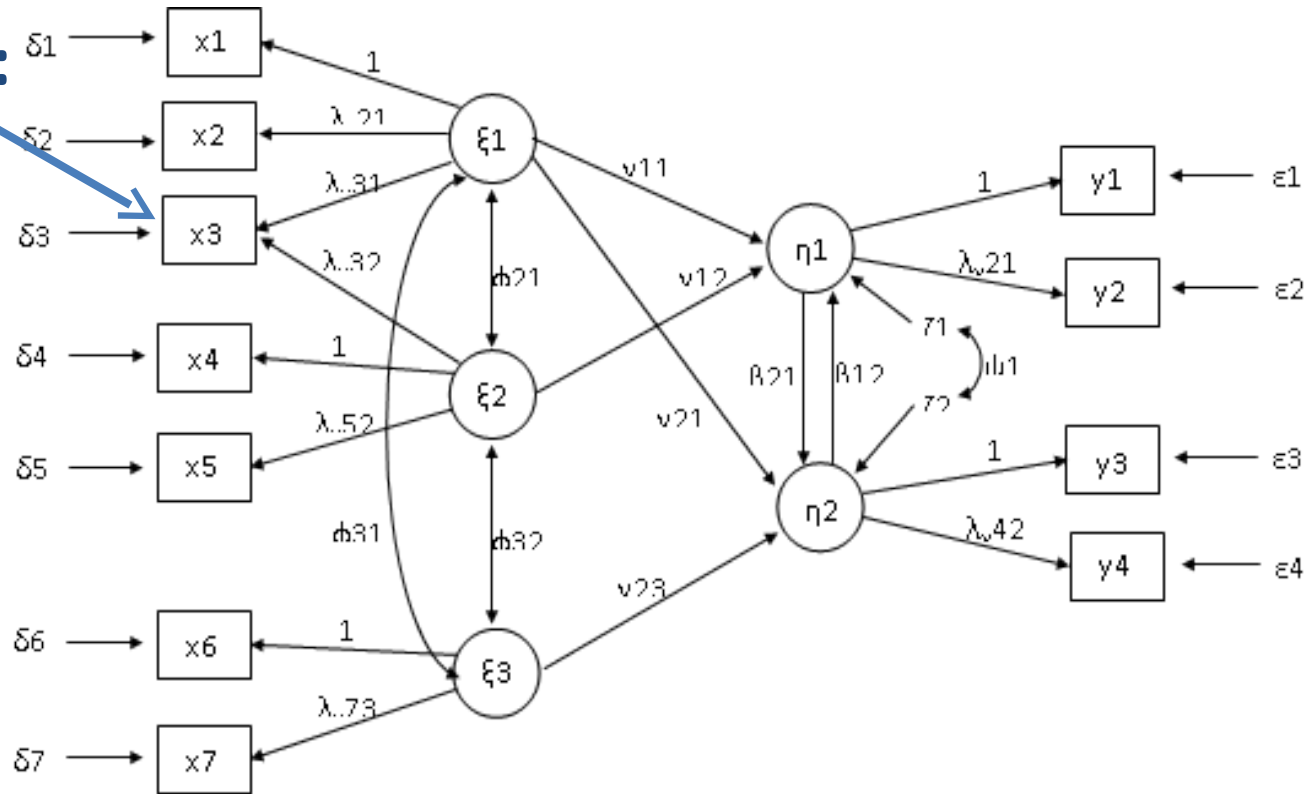


Full SEM model



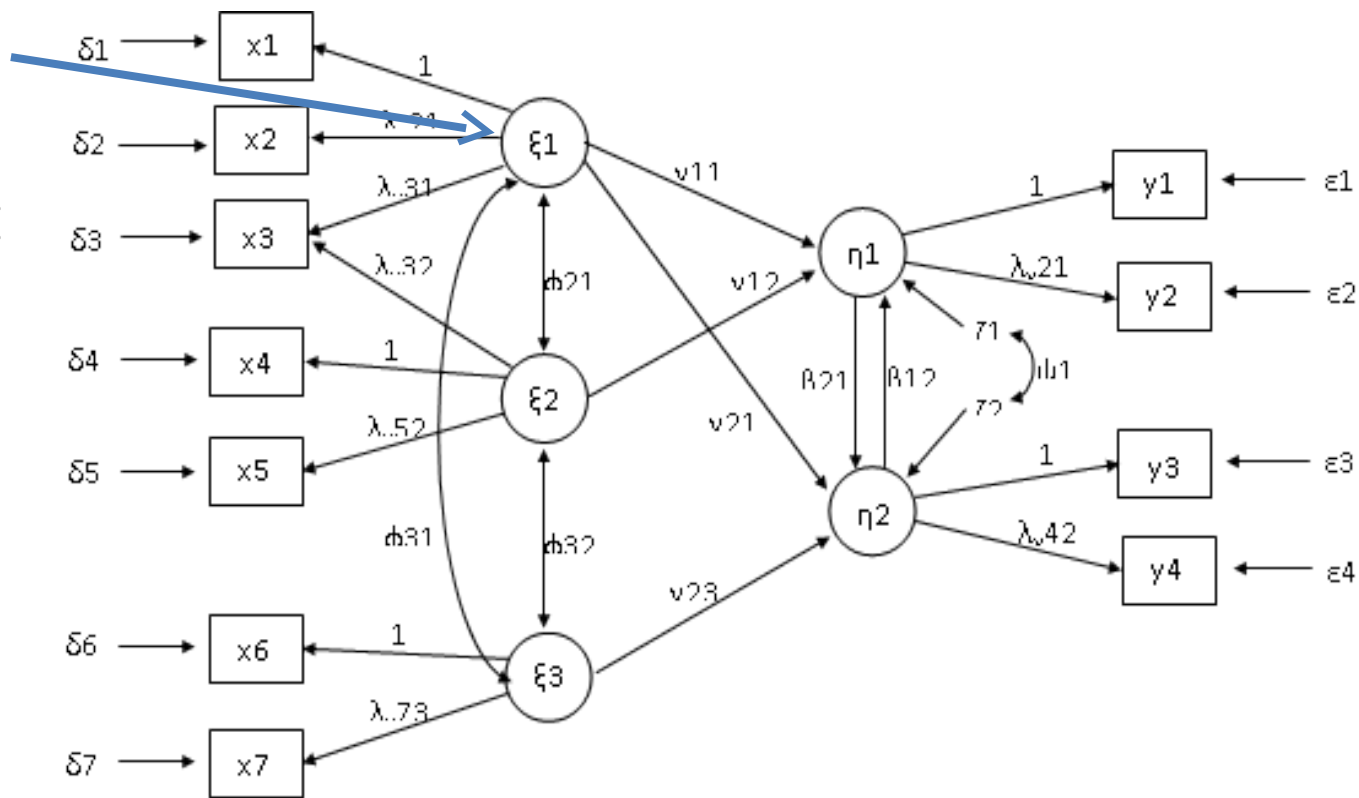
Key terms

Manifest variable:
variable in the
model that is
measured directly
(represented by
square)



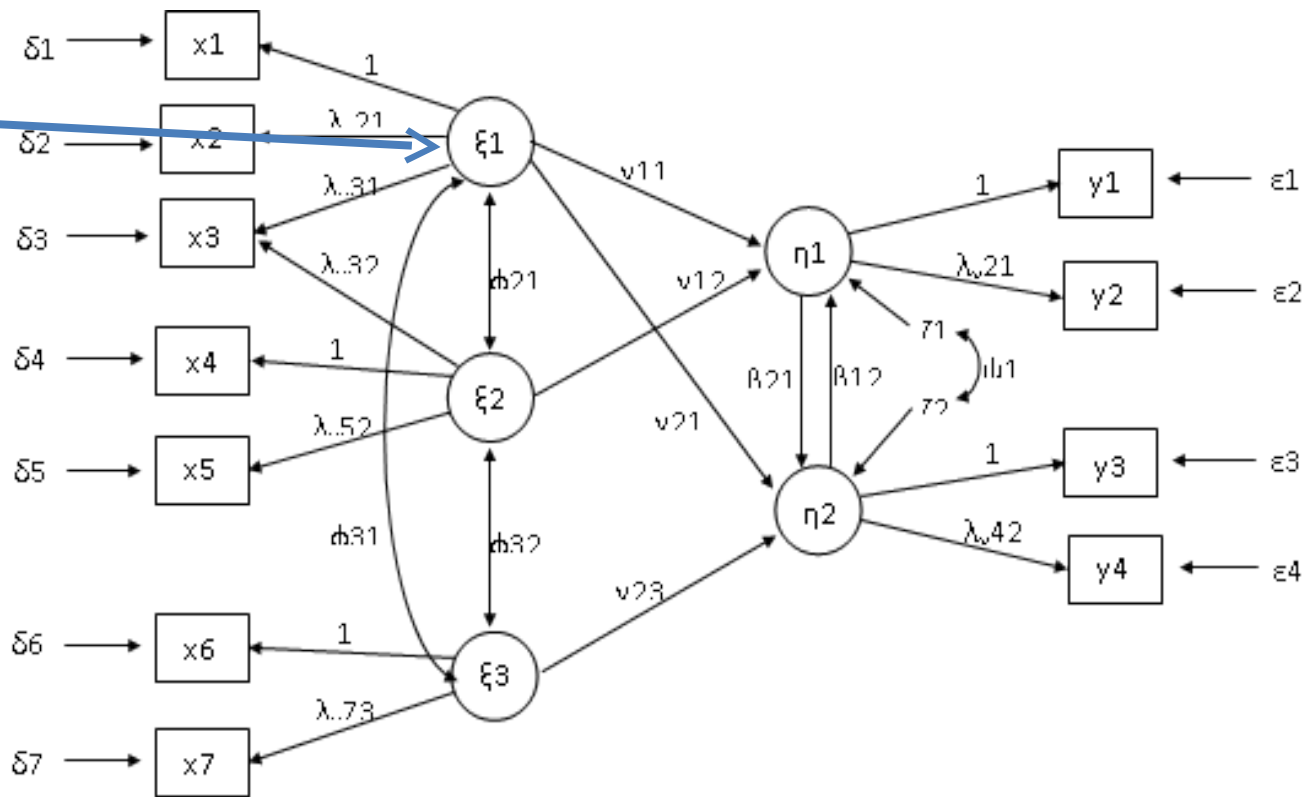
Key terms

Latent variable: variable in the model that is not measured directly, but estimated based on set of manifest variables (represented by ovals)



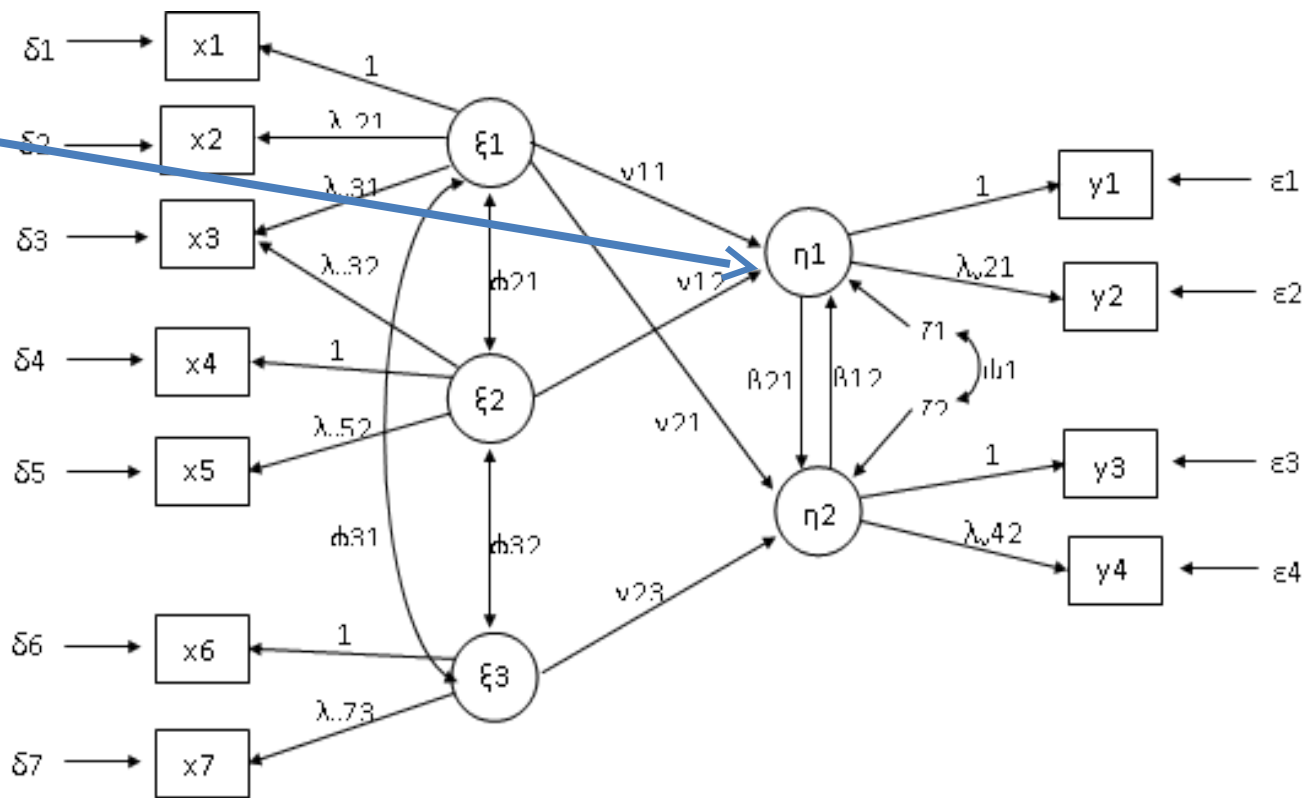
Key terms

- **Exogenous variable:** variable that is not caused by another, but usually causes one or more variables in the model



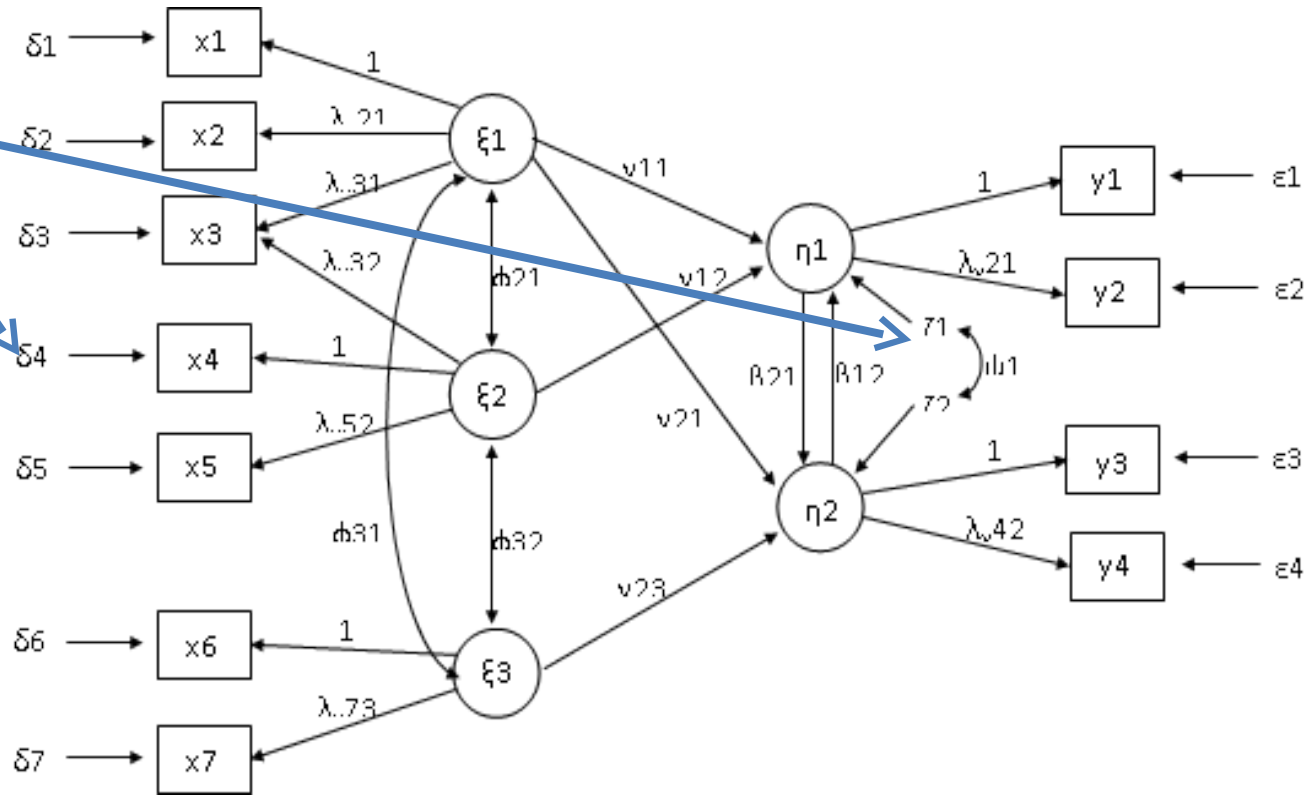
Key terms

- **Endogenous variable:** variable that is caused by another, but can also cause one or more variables in the model



Key terms

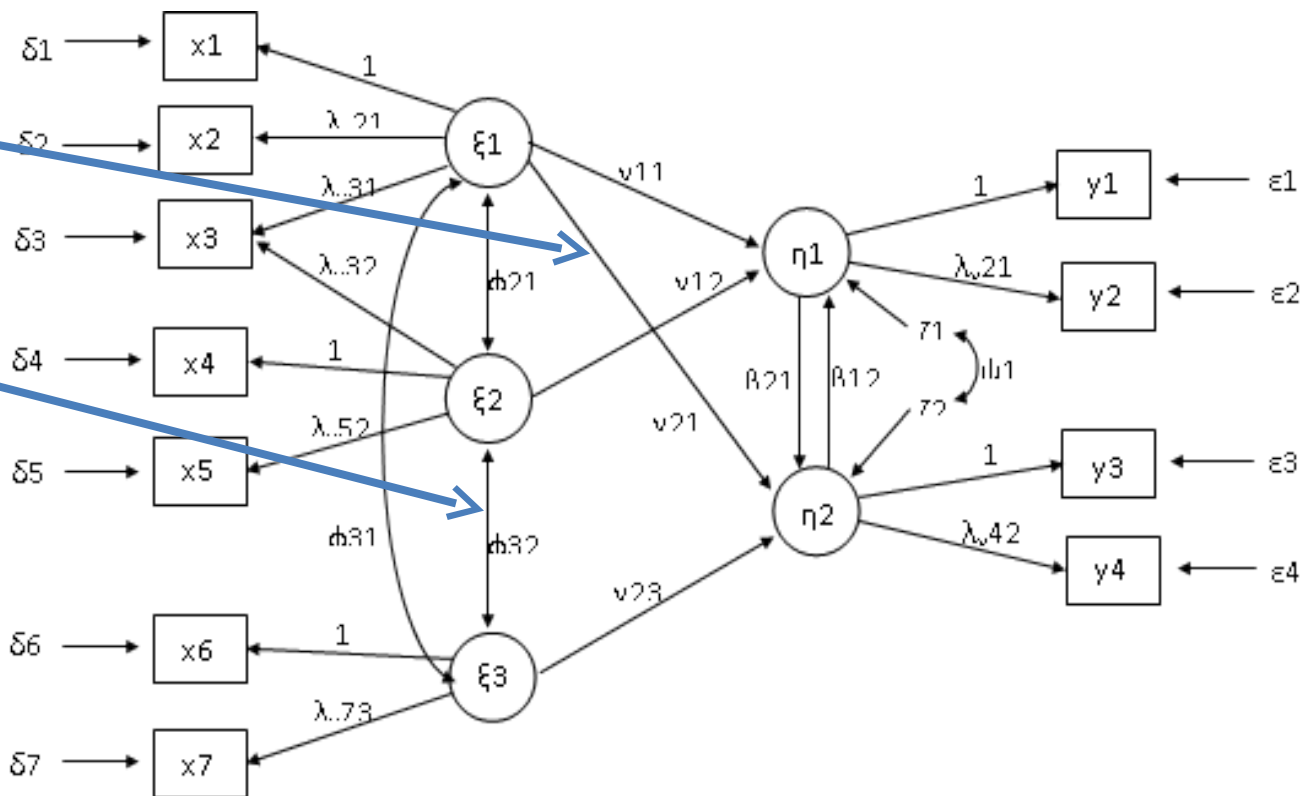
- Error variance:** variance in endogenous variable or manifest variable that cannot be explained by variable in the model



Key terms

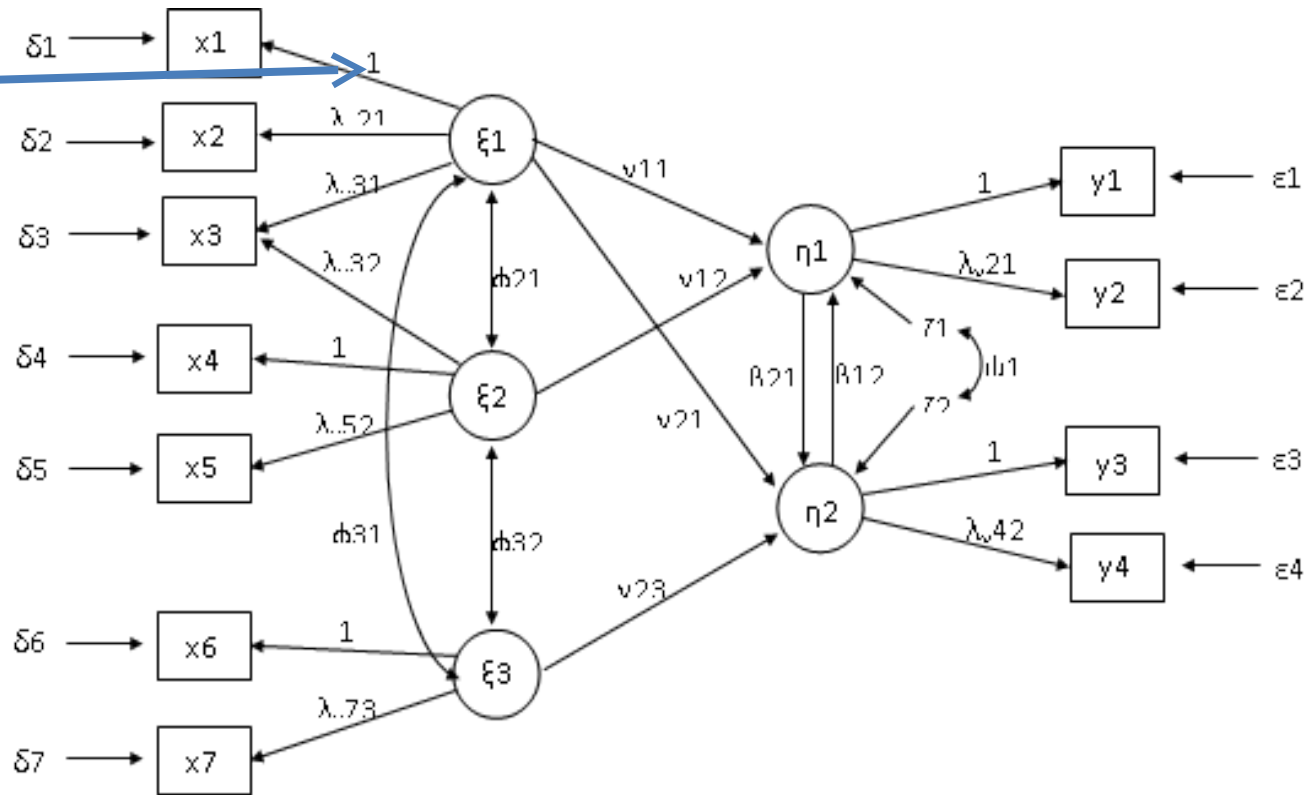
- Causal relation

- Covariance relation



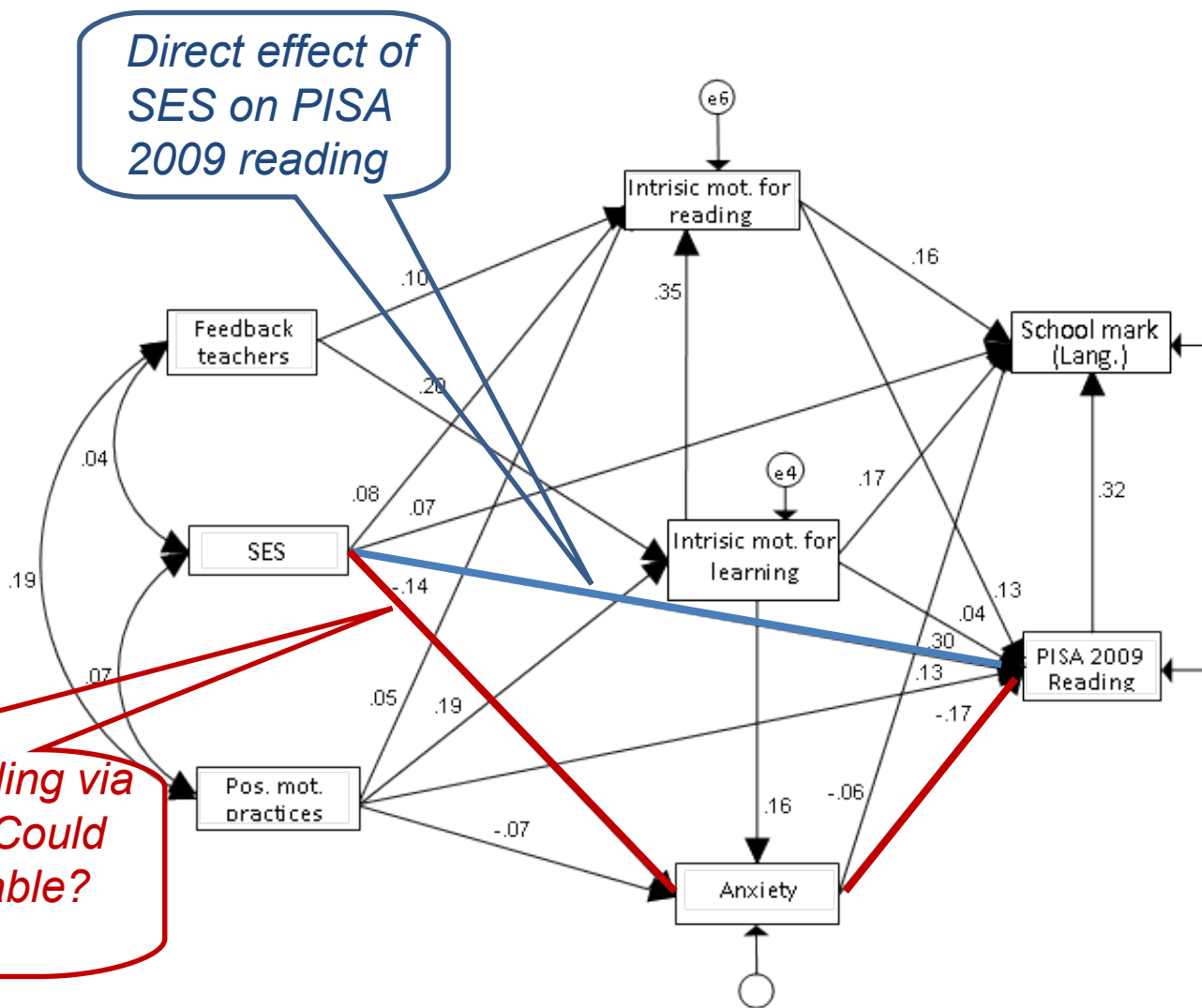
Key terms

- Define the scale for latent variable: since latent variables are estimated they do not have their scale. Because of that regression between latent variable and one manifest variable should be fixed at 1



Key terms

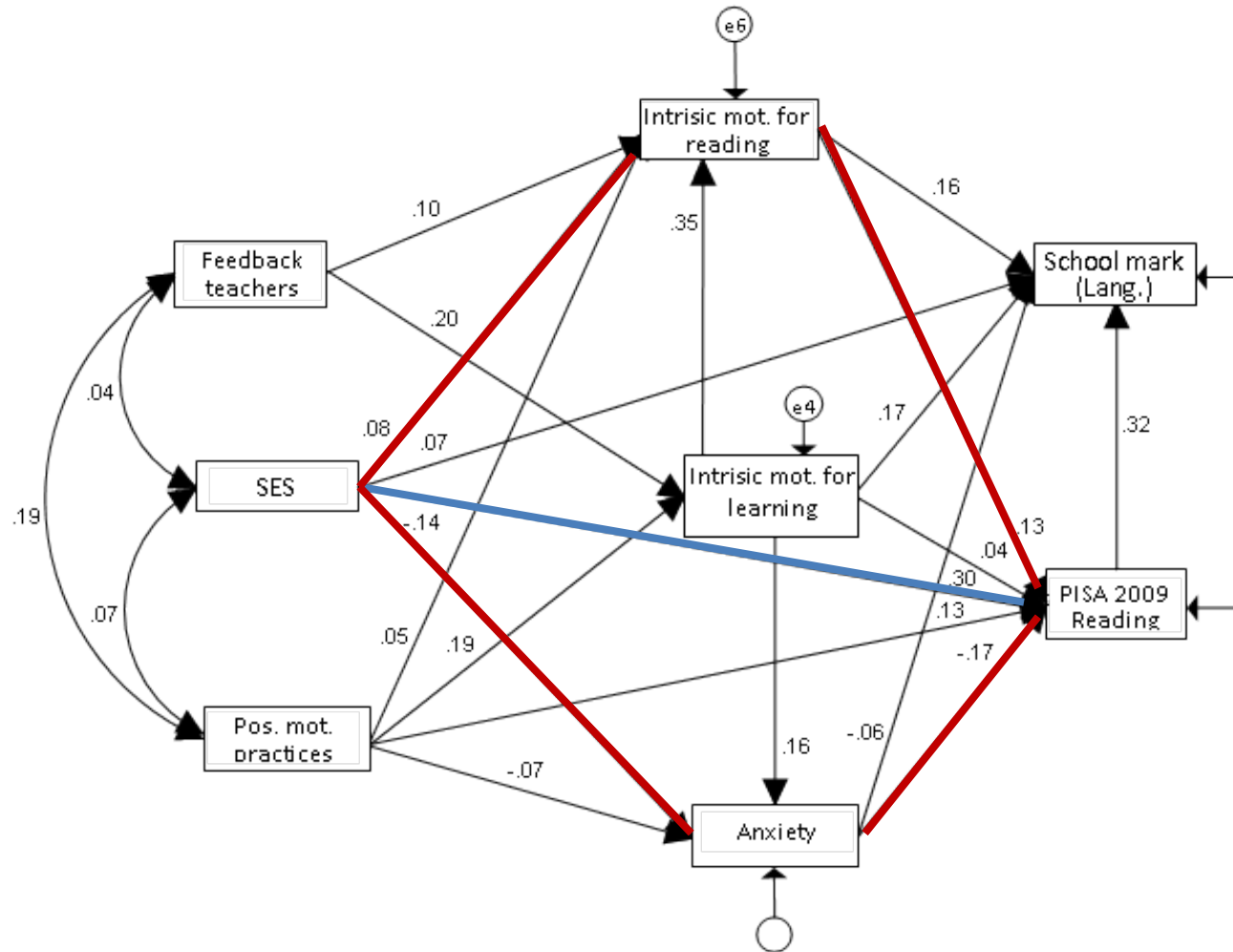
- **Direct effect:** direct causal effect of one variable on the other
- **Indirect effect:** when variable has effect on other via third variable (mediation)
- **Total effect:** sum of direct and indirect effects



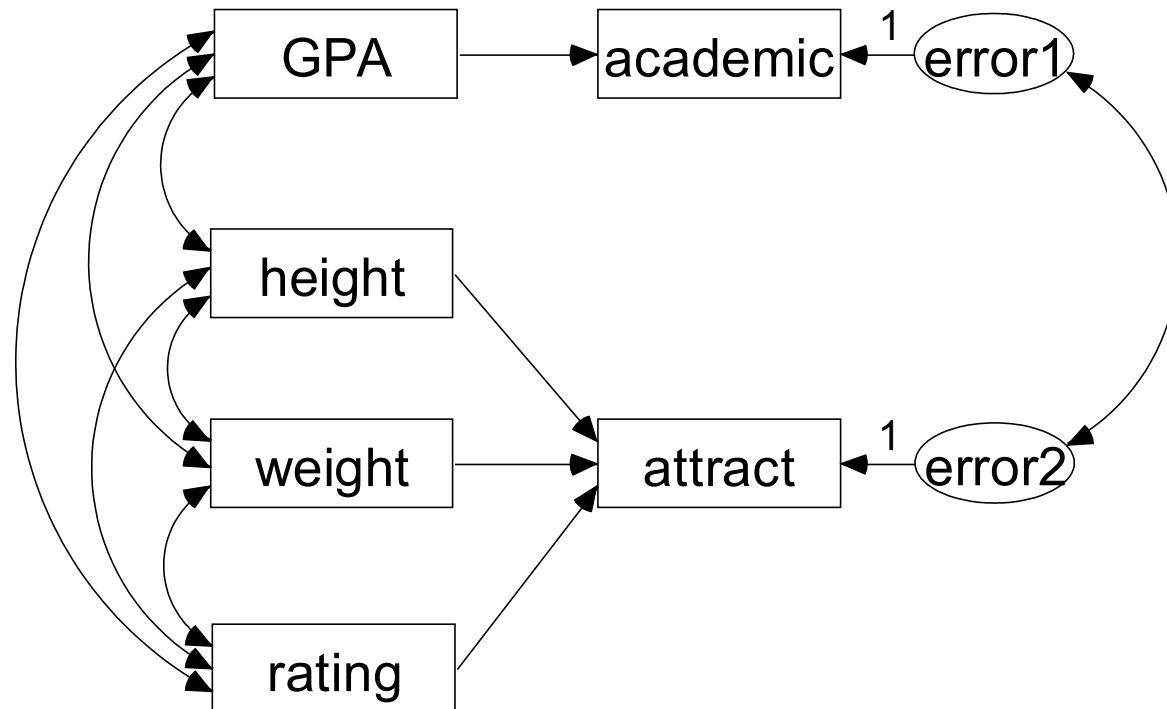
Indirect effect of SES on reading via Anxiety and some other var. Could you find other mediating variable? (see next page for answers)

Correct answers from previous slide

- **Direct effect:**
- **Indirect effects:**



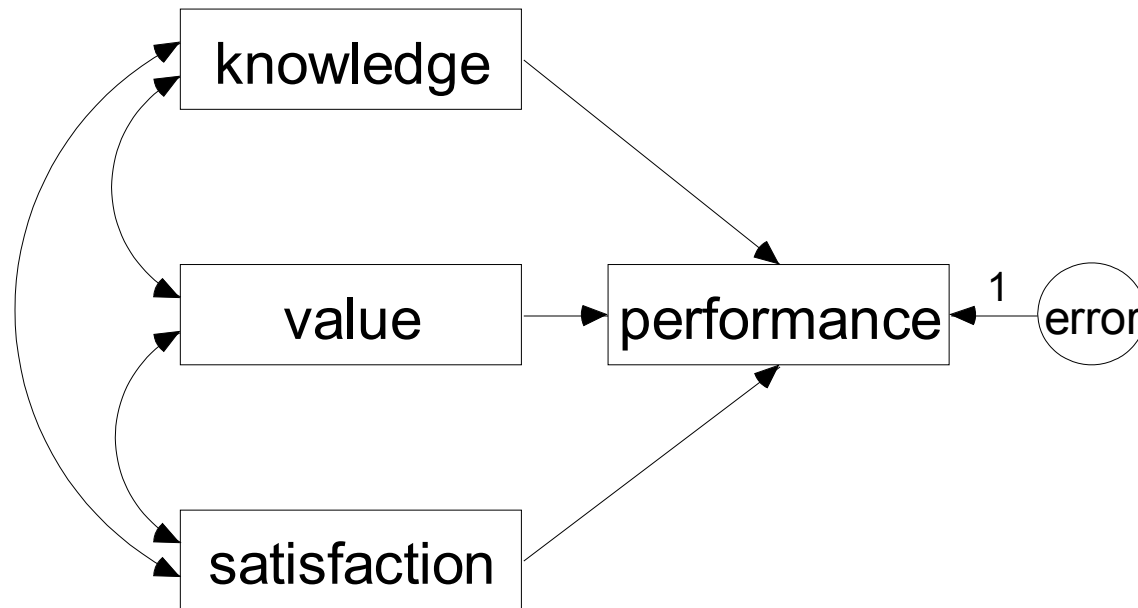
Reduced model: path model (only structural model)



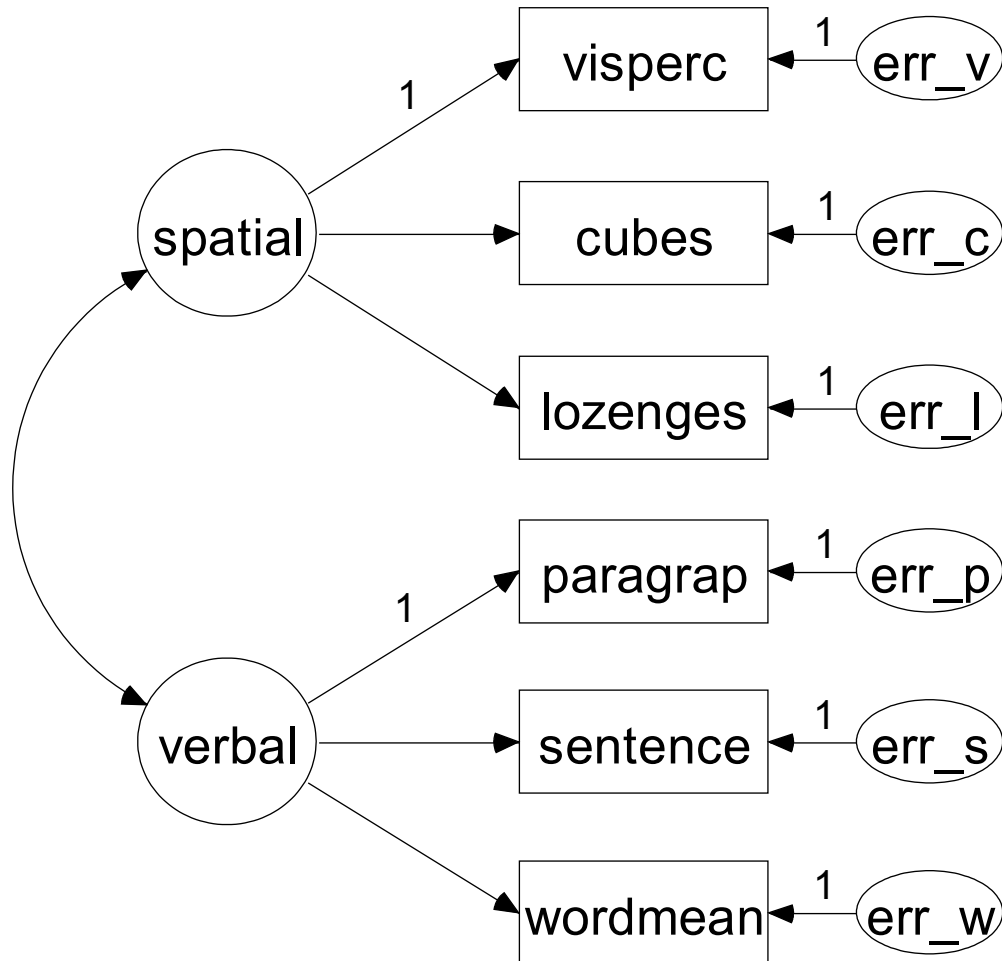
SEM analysis as general model

- Some frequently used multivariate kinds of analysis (e.g. regression, and factor analysis) can be treated as special cases of the SEM analysis

Multiple regression analysis



Confirmatory factor analysis



How to use the SEM?

- Main purpose of the SEM analysis is to confirm some model that is defined based on certain theory or previous research studies
 - Comparison of different models
 - For example, two theories assume different relationships within same set of variables. You can define two models based on these two theories and to check by the SEM analysis which one fits better to your data
 - Comparison of same model in different subgroups
 - For example, a theory assume some relationships between some variables. You wish to test whether that model is valid for both males and females (or for participants from different countries)

Whether you can use SEM for exploratory purposes?

- Sometimes you have not enough ground to define model. In this situation you might be tempted to use SEM to explore different model and to find the one with best fit.
- Many authors discourage such use of the SEM
- But, the SEM is used in this way as well.
- However, you should be careful because data driven models/theories can be misleading

Key steps in the SEM analysis

- To define model (based on theory or previous research studies)
- Software search for set of free parameters that will minimize difference between reproduced covariance and covariance calculated based on empirical data
- Check validity of the model
- If the model fit to data then each parameter in the model is analyzed

Unfortunately, SEM asks for big sample

- For small models (few variables): Min. 150-200 of participants per group
- For other models: Min. 5 times more than number of manifest variables

Some basic fit measures for model

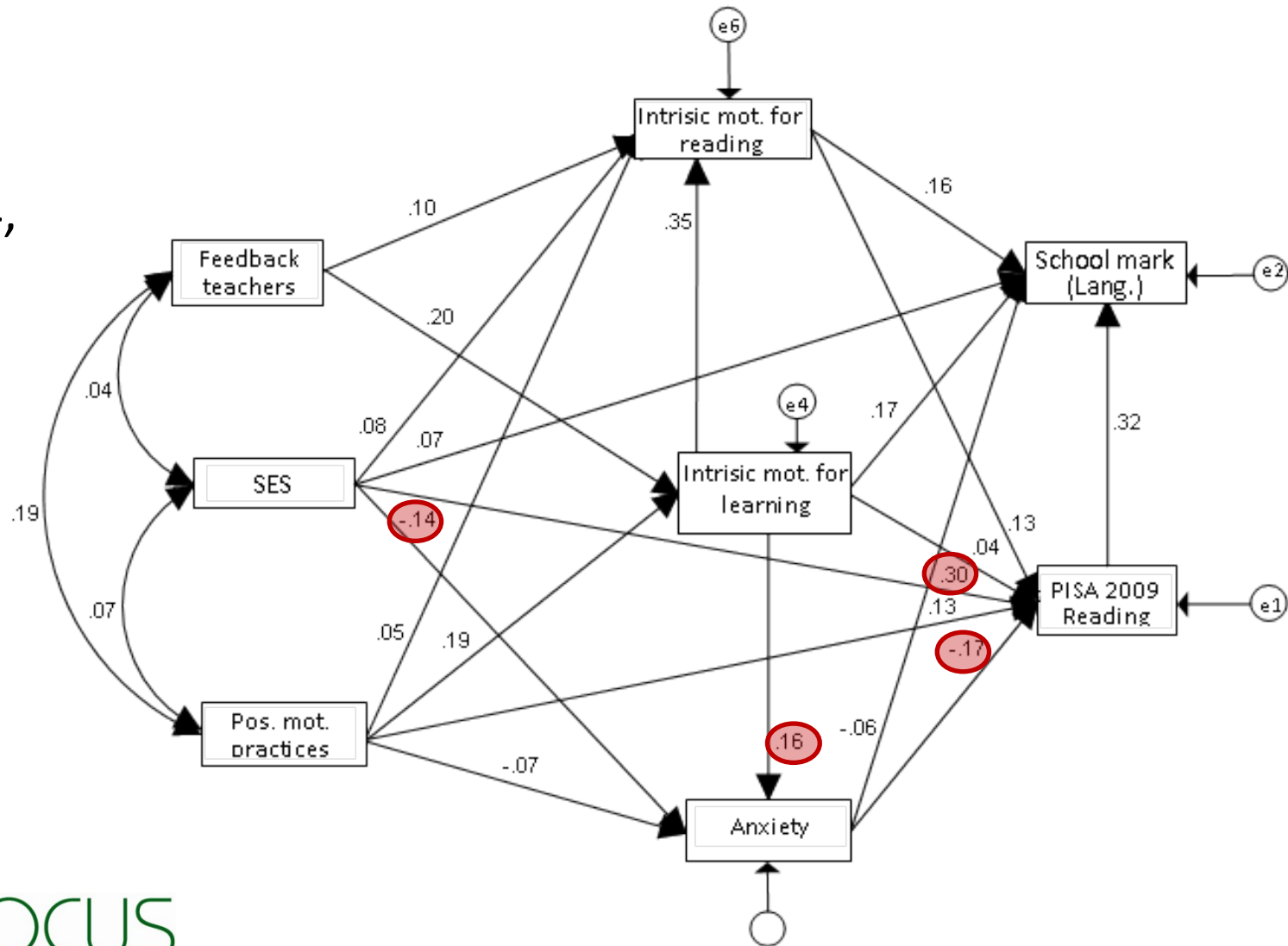
- χ^2 - if it is significant it means that model DOES NOT fit properly data. However, it is very sensitive and in many case suggest that model does not fit although there is no big difference between model and data
- $\chi^2/df < 2-3$ (good) < 5 (acceptable)
- RMR (root mean square residual) – it should be less than .05 (if it is less than .10 the model asks for further research)
- RMSEA (root mean square error of approximation)– similar to RMR except there is statistical test (p-close) showing whether it is less than .05 at the population level

Analysis at the level of parameter

- For each parameter it is possible to test two hypothesis:
 - Whether it is different than zero
 - Whether it is different than some other parameter

A SEM example: development of reading competence in school

- Model fits data
Chi(6)=10.34,
p=.11
- It reveals several very interesting findings about direct and indirect effects on reading competence



Programi

- LISREL (prvo SPSS, sad samostalan program)
www.ssicentral.com/lisrel
- **AMOS (prvo samostalan, sad SPSS)**
www.spss.com/amos
- EQS www.mvsoft.com/products.htm
- RAMONA (Systat) www.systat.com
- CALIS (SAS)
v8doc.sas.com/sashtml/stat/chap19/sect5.htm

Group work

- Divide yourself in smaller groups
- Each group will focus on one type of food
- Discuss and define a SEM model based on FOCUS dataset that would be worth of validate
- Formulate justification for model – what is theoretical and empirical background supporting such model

Thank you for your attention.

Have a nice scaling and modelling in
future 😊