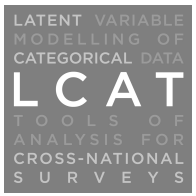
Latent Trait and Latent Class Analysis for Multiple Groups

Day 2: Multigroup analysis

LCAT Training Workshop

2012



Outline of the workshop

- Day 1: Models for single groups
  - Session 1.1: Introduction and latent trait models
  - Session 1.2: Latent class models and model assessment
- Day 2: Models for multiple groups
  - Session 2.1: Cross-group comparisons of latent distributions
  - Session 2.2: Examining measurement equivalence and non-equivalence
- Each session consists of a lecture and a computer class

Session 2.1

Structural models for multigroup models

Outline of Session 2.1

- Multigroup modelling: Introduction
  - For example: Analysis of cross-national surveys
- Structural models with 1 latent variable: Cross-group comparisons of latent distributions
- Structural models with 2 latent traits

Latent variable models with covariates

Return to the general formulation of a latent variable model:

$$p(\mathbf{y}, \boldsymbol{\eta}|\mathbf{x}) = p(\mathbf{y}|\boldsymbol{\eta}, \mathbf{x})\, p(\boldsymbol{\eta}|\mathbf{x})$$

where

- $p(\mathbf{y}|\boldsymbol{\eta}, \mathbf{x})$ is the measurement model
- $p(\boldsymbol{\eta}|\mathbf{x})$ is the structural model

and $\mathbf{x}$ are observed explanatory variables ("covariates")

Latent variable models with covariates

The import of covariates $\mathbf{x}$ depends on where they appear:

- If the structural model $p(\boldsymbol{\eta}|\mathbf{x})$ depends on $\mathbf{x}$, the distribution of latent variables $\boldsymbol{\eta}$ depends on the covariates. This is usually substantively interesting.
- If the measurement model $p(\mathbf{y}|\boldsymbol{\eta}, \mathbf{x})$ depends on $\mathbf{x}$, the response probabilities of some items in $\mathbf{y}$ depend on the covariates, even given the latent variables $\boldsymbol{\eta}$. There is then **non-equivalence of measurement**. This is usually a nuisance.

In this session, we

- assume $p(\mathbf{y}|\boldsymbol{\eta}, \mathbf{x}) = p(\mathbf{y}|\boldsymbol{\eta})$, i.e. measurement equivalence. This will be changed in the last session.
- focus on the structural models $p(\boldsymbol{\eta}|\mathbf{x})$, and what dependence of covariates means there.

Cross-national social surveys

- Primary motivation of the LCAT project
- Surveys where the same questions (translated) are asked in several countries
- For example:
  - Eurobarometers (and other regional "barometer" surveys, for Africa, Asia and Latin America)
  - World Values Survey
  - European Values Study
  - International Social Survey Programme
  - European Social Survey
- A key purpose is obviously answering comparative cross-national research questions.

Multigroup analysis of cross-national data

Suppose that we have data on respondents from countries $g = 1, \ldots, G$.

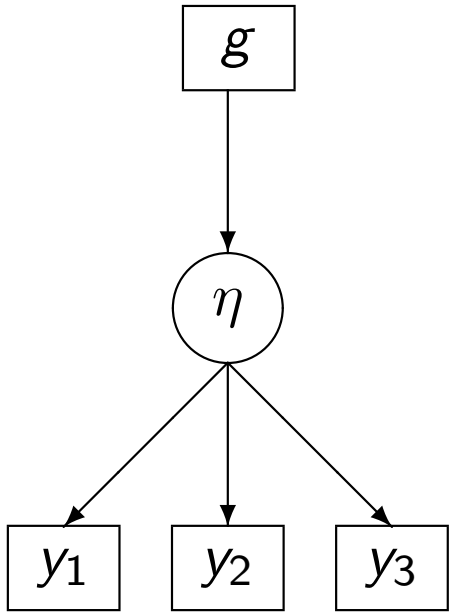
A respondent's country can be used as an explanatory variable.

- Take $\mathbf{x} = (x_2, \ldots, x_G)$, where x_g is a dummy variable for country g.

Exactly the same ideas apply if instead of countries we compare any distinct **groups**, such as different regions, men vs. women, different rounds (years) of a repeated survey, etc.

Multigroup latent variable models

- Path diagram of the multigroup latent variable model (here with one η) is thus of the following form:
- Comparative research questions are usually about the relationship between group g and latent variables $\boldsymbol{\eta}$.
- In other words, how do distributions of **individual-level** latent variables $\boldsymbol{\eta}$ vary across groups?
- In this session we focus on how this relationship might be modelled.



More general models for cross-national analysis

We focus on cases where the country (group) is the only explanatory variable. This leads to comparisons of averages, variances and proportions between countries, in a fairly descriptive spirit.

More general versions are also possible, for example:

- Instead of country dummies, using one or more country-level variables (GDP, type of welfare state, etc.) as explanatory variables.
  - Seems more focused, especially if choice of these variables is theoretically motivated.
  - However, we do not have many theories about how specific country-level characteristics affect individual-level constructs.
  - If not theoretically motivated, any such country-level variables may be essentially proxies for the country dummies. This is especially so with smallish numbers of groups, such as 20–40 in European surveys.

More general models for cross-national analysis

- Including also (or only) individual-level covariates (e.g. age, sex, education etc.).
  - This allows us to consider both individual-level (micro) associations between covariates and latent variables, and how these may vary across countries (micro-macro interactions).
  - This is straightforward in Mplus, but not part of this workshop.
- Instead of country dummies ("fixed effects"), use random effects for countries.
  - This defines **multilevel models** for latent response variables η.
  - Can be done, but does not necessarily add any value, especially when the number of countries is small.

Multigroup notation

When group is the only covariate, we can also omit $\mathbf{x}$ from the notation and write instead

$$p^{(g)}(\mathbf{y}, \boldsymbol{\eta}) = p^{(g)}(\mathbf{y}|\boldsymbol{\eta})\, p^{(g)}(\boldsymbol{\eta})$$

where supercript (g) indicates that a distribution varies across groups $g = 1, \ldots, G$.

For the moment we assume $p^{(g)}(\mathbf{y}|\boldsymbol{\eta}) = p(\mathbf{y}|\boldsymbol{\eta})$, i.e. measurement equivalence across the groups.

- The measurement models are then just like in the single-group case yesterday.

Focus now on the structural model $p^{(g)}(\boldsymbol{\eta})$, starting with the one-$\eta$ case.

Structural models with 1 latent variable

How does $p^{(g)}(\eta)$ depend on group $g = 1, \ldots, G$?

Latent trait models:

$$\eta \sim N(\kappa^{(g)}, \phi^{(g)})$$

where the parameters need to be constrained in one group, say $(\kappa^{(1)}, \phi^{(1)}) = (0, 1)$.

Latent class models:

$$\alpha_c^{(g)} = P^{(g)}(\eta = c)$$

Both of these can also be formulated as regression models for η given group dummies $\mathbf{x} = (x_2, \ldots, x_G)$.

- Linear model for a latent trait, multinomial logistic for a latent class.

Basic output of the analysis is then a table of estimates of these parameters across groups.

Cross-group equalities

We may also examine hypotheses that the distribution of the latent variable is the same across groups:

- $(\kappa^{(g)}, \phi^{(g)}) = (\kappa, \phi) = (0, 1)$ for all g (or perhaps just $\phi^{(g)} = \phi = 1$)
- $\alpha_c^{(g)} = \alpha_c$ for all g for every c

These can be tested with a likelihood ratio test against a model where the parameters are not constrained to be equal.

We might also consider hypotheses where parameters are equal within some subsets of groups, at least if there is a theoretical motivation for such subgrouping.

Fitting multigroup models in Mplus

- For both latent class and latent trait models, same model can be fitted in two different-looking ways:
  1. "Covariate specification" which uses group dummies **x** explicitly as explanatory variables.
  2. "Multiple-group specification" which does not.
- Here we use one of each (1. for latent classes, 2. for latent traits).
  - See `stats.lse.ac.uk/lcat` for the other way.
- There are some differences between them in speed, convenience and which models can be fitted.
  - In the free demo version of Mplus, the covariate specification allows only three groups (two dummies).

Example: Latent class model with 2 groups

Data from Round 5 of European Social Survey (2010).

"To what extent do you think [country] should allow people [type] to come and live here?"

with three questions on different [types] of people:

- **same**: *...of the same race or ethnic group as most [country]'s people..."*
- **differ**: *...of a different race or ethnic group from most [country] people..."*
- **poor**: *...from the poorer countries outside Europe..."*

Response options for each: Allow (1) *Many*, (2) *Some*, (3) *A few*, (4) *None*.

Compare data for two countries, **Finland** and **Sweden**.

Latent class model with 2 groups: Mplus input

```
Title:
  Attitudes to immigration, ESS5.
  4-class latent class model.
  2 groups: Finland and Sweden.
Data:
  File =immigr.dat ;
Variable:
  Names = same differ poor gr1-gr20 country;
  Missing = all (99) ;
    Useobservations = (gr19==1 OR gr9==1);
        ! Here gr19 is dummy for Sweden, gr9 for Finland.
    Nominal = same differ poor;
    Usevariables = same differ poor gr19;
    Classes = class(4);
Analysis:
    Type=Mixture;
    Estimator=ML;
    Starts=30 10;
Model:
    %overall%
        class ON gr19;
!        class ON gr19@0; ! Use this to make class independent of country
Savedata:
    File="tmp.dat";
```

Immigration example: Item probabilities

A 4-class model fits the data reasonably well:

		Latent class			
Item	Level	'Many'	'Some'	'A few'	'None'
same	many	.98	.08	.05	.05
	some	.02	.91	.30	.05
	a few	.00	.01	.65	.47
	none	.00	.00	.00	.43
differ	many	.99	.01	.00	.01
	some	.01	.96	.02	.02
	a few	.00	.03	.94	.00
	none	.00	.00	.03	.97
poor	many	.87	.03	.00	.00
	some	.11	.80	.03	.04
	a few	.02	.16	.82	.08
	none	.00	.00	.14	.88

Clearly most respondents do not distinguish much between different types of potential immigrants.

Immigration example: Probabilities of latent classes

Country	Probability of latent class			
	'Many'	'Some'	'A few'	'None'
Finland	.06	.31	.55	.08
Sweden	.35	.55	.09	.01

(Larger of the two probabilities for each class in gray.)

Differences between the two countries are clearly large, and also statistically significant:

```
> lcat.lrtest(workshop.imm,1,2)

Likelihood ratio test:
H0: imm4cl_indep        H1: imm4cl
LR = 1147.81    df = 3  P-value = <0.0005
```

Example: Latent trait model with 23 groups

Data from Round 3 of European Social Survey (2006), "Personal and social well-being" rotating module.

- **optim**: *Im always optimistic about my future.*
- **positive**: *In general I feel very positive about myself.*
- **failure**: *At times I feel as if I am a failure.*
- **ideal**: *On the whole my life is close to how I would like it to be*

Response options for each: Agree strongly, Agree, Neither agree nor disagree, Disagree, Disagree strongly.

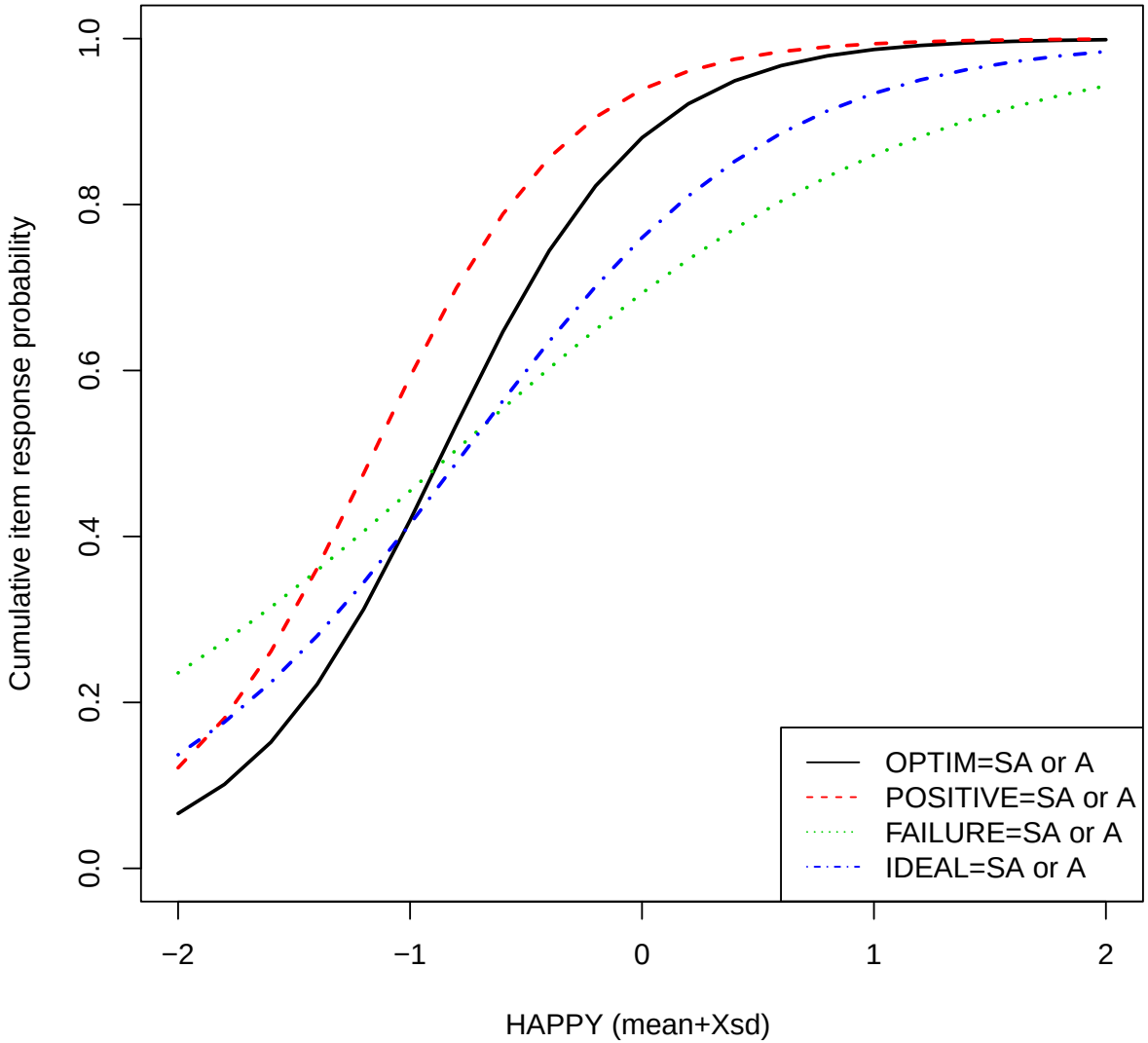
Use data for all 23 countries in round 3 of ESS.

Fit a 1-trait model, with the items treated as ordinal.

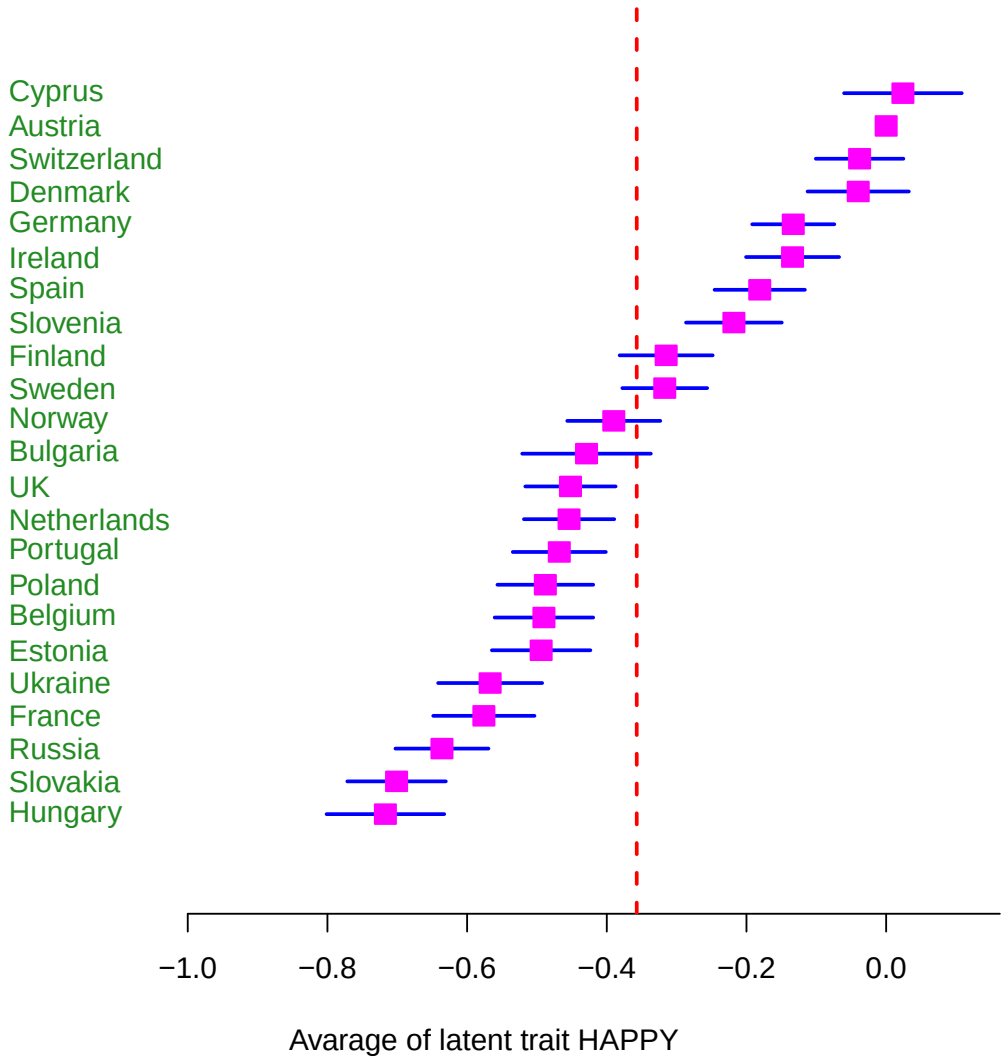
Latent trait model with 23 groups: Mplus input

```
Title:
  Subjective well-being, ESS3, 23 countries.
  1-trait ordinal latent trait model.
Data:
  File = ess3happy.dat ;
Variable:
  Names = optim positive failure ideal gr1-gr23 cntry;
  Missing = all (99) ;
    Usevariables = optim positive failure ideal;
    Categorical = optim positive failure ideal;
    Classes = country(23);
    Knownclass= country (
        cntry=1 cntry=2 cntry=3 cntry=4 cntry=5 cntry=6 cntry=7 cntry=8
        cntry=9 cntry=10 cntry=11 cntry=12 cntry=13 cntry=14 cntry=15 cntry=16
        cntry=17 cntry=18 cntry=19 cntry=20 cntry=21 cntry=22 cntry=23);
Analysis:
    Type = Mixture;
    Estimator=ML; Algorithm = Integration; Starts = 20 10;
Model:
    %overall%
        happy BY optim* positive failure ideal;
        [happy@0]; happy@1;
    %country#2%
        [happy]; happy;
... and so on ...
    %country#23%
        [happy]; happy;
Savedata:
    File="tmp.dat"; Save=Cprobabilities;
```

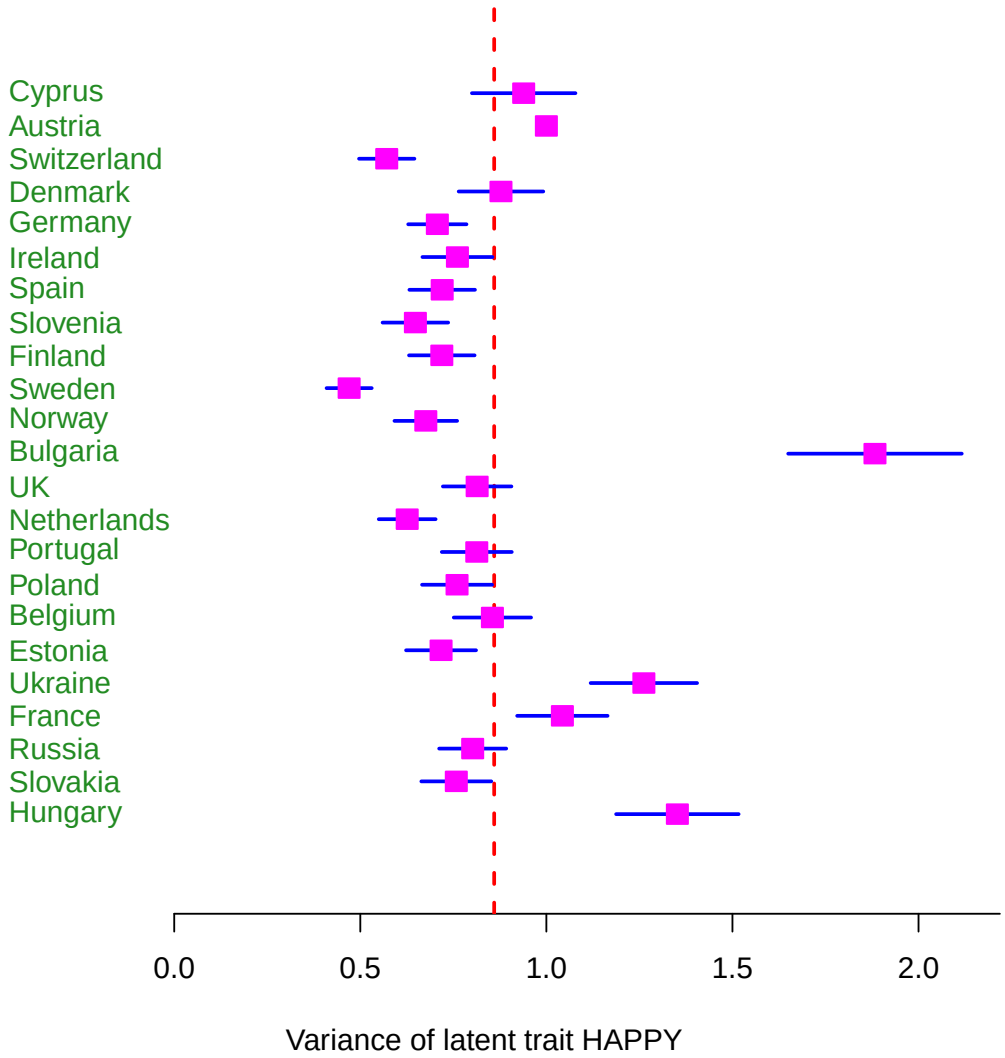
Happiness example: Item response curves



Happiness example: Estimates of national averages



Happiness example: Estimates of national variances



Models with 2 (or more) latent traits

(Analogous latent class models are also possible, but less common and not considered here.)

Consider 2 latent traits $\boldsymbol{\eta} = (\eta_1, \eta_2)$ and suppose that η_1 is treated as an explanatory variable for η_2.

Suppose that in group g, we have $\eta_1 \sim N(\kappa_1^{(g)}, \phi_{11}^{(g)})$ and

$$\eta_2 = \gamma_0^{(g)} + \gamma_1^{(g)} \eta_1 + \zeta \quad \text{with} \quad \zeta \sim N(0, \psi^{(g)}).$$

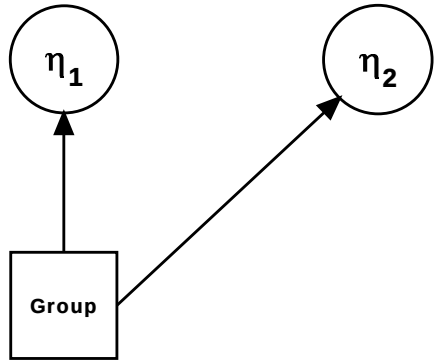
Even with just two latent variables, various possibilities arise here, depending on which (if any) parameters are equal across groups.

Nested models here can be compared using likelihood ratio tests.

2 traits: Some structural models

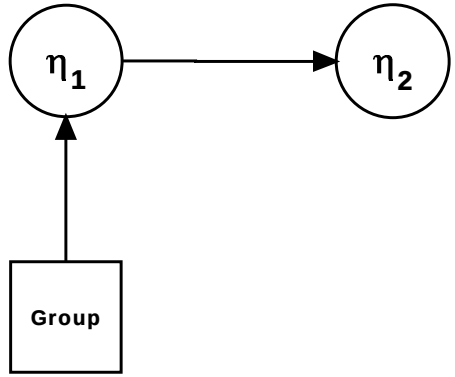
η_1 and η_2 are independent given group:

$$\eta_2 = \gamma_0^{(g)} + \zeta$$



η_2 is independent of group given η_1:

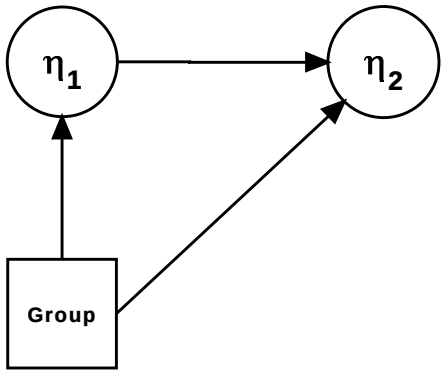
$$\eta_2 = \gamma_0 + \gamma_1 \eta_1 + \zeta$$



2 traits: Some structural models

Association of η_1 and η_2 is the same across groups:

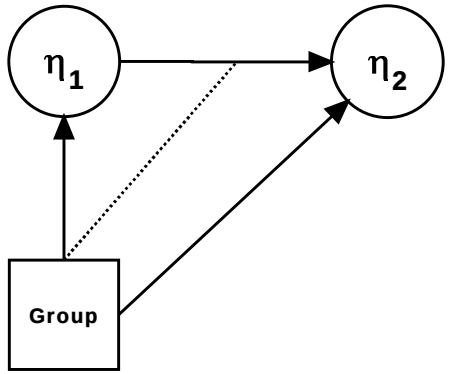
$$\eta_2 = \gamma_0^{(g)} + \gamma_1 \eta_1 + \zeta$$



Everything in the model depends on group also:

$$\eta_2 = \gamma_0^{(g)} + \gamma_1^{(g)} \eta_1 + \zeta$$

(i.e. there is an interaction between group and η_1 in the model for η_2)



Example: 2-trait model with 2 groups

Data from Round 4 of European Social Survey (2008), "Welfare attitudes in a changing Europe" rotating module.

Three items on *perceived risk*:

How likely is it that during the next 12 months...

- **unemp12**: *...you will be unemployed and looking for work for at least four consecutive weeks?* (Include only respondents who answered this.)
- **care12**: *...you will have to spend less time in paid work than you would like, because you have to take care of family members or relatives?*
- **money12**: *there will be some periods when you dont have enough money to cover your household necessities?*

Originally 4 response options for each, here collapsed into two as "Not at all likely/Not very likely" vs. "Likely/Very likely".

Example: 2-trait model with 3 groups

Three items on *attitudes toward welfare recipients*:

- **nottry**: *Most unemployed people do not really try to find a job.*
- **notent**: *Many people manage to obtain benefits and services to which they are not entitled.*
- **pretend**: *Employees often pretend they are sick in order to stay at home.*

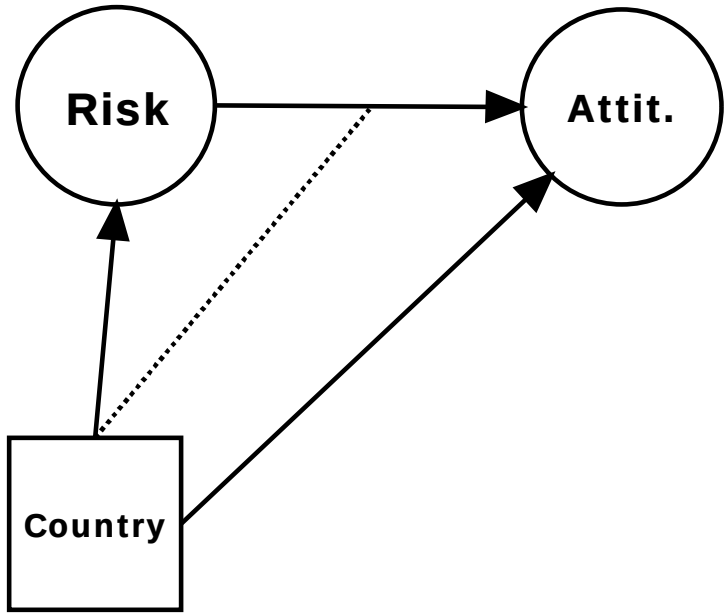
Originally 5 response options for each, here collapsed into three as "Agree strongly/Agree", "Neither agree nor disagree" and "Disagree strongly/Disagree".

Use data for 3 countries: UK, Germany and Sweden.

Ordinal measurement model for each set of items, with no cross-loadings.

Example: 2-trait model with 3 groups

Consider a model for *Attitude* (η_2) given *Risk* (η_1) and country:



2-trait model 3 groups: Mplus input (part only)

```
Model:
    %overall%
        risk BY unemp12* care12 money12;
        attitude BY nottry* notent pretend;
        attitude ON risk;
        [risk@0]; risk@1; [attitude@0]; attitude@1;
    %country#2%
        attitude ON risk;
        [risk]; risk; [attitude]; attitude;
    %country#3%
        attitude ON risk;
        [risk]; risk; [attitude]; attitude;
```

- Other parts are very similar to the 1-trait example above.
- Here parameters vary across countries if they are repeated under each country-specific model statement:
  - Risk (η_1): mean $\kappa_1^{(g)}$ is `[risk]` and variance $\phi_{11}^{(g)}$ is `risk`.
  - Attitude (η_2): Intercept $\gamma_0^{(g)}$ is `[attitude]`, regression coefficient $\gamma_1^{(g)}$ is `attitude ON risk` and residual variance $\psi^{(g)}$ is `attitude`.

Welfare example: Estimated structural model

$$\eta_1 \sim N(\kappa_1^{(g)}, \phi_{11}^{(g)}) \quad \text{and} \quad \eta_2 = \gamma_0^{(g)} + \gamma_1^{(g)}\eta_1 + \zeta, \quad \zeta \sim N(0, \psi^{(g)})$$

Where η_1 is perceived risk (high values = high risk) and η_2 attitude toward welfare recipients (high values = positive attitude).

	Risk			Attitude				
Country	$\kappa_1^{(g)}$	(s.e.)	$\sqrt{\phi_{11}^{(g)}}$	$\gamma_0^{(g)}$	(s.e.)	$\gamma_1^{(g)}$	(s.e.)	$\sqrt{\psi^{(g)}}$
UK	0		1	0		−0.08	(0.04)	1
Germany	−0.12	(0.09)	0.97	1.22	(0.06)	−0.19	(0.05)	1.07
Sweden	−0.24	(0.09)	0.66	1.60	(0.09)	−0.36	(0.12)	1.26

Note that the estimated association $\hat{\gamma}_1^{(g)}$ is negative. Variation in this between the countries is only mildly significant ($p = 0.02$). If it is constrained to be equal across countries, we get $\hat{\gamma}_1 = -0.14$.

Session 2.2

Measurement equivalence and nonequivalence in multigroup models

Outline of Session 2.2

- Measurement equivalence and nonequivalence
- Modelling nonequivalence
- Model assessment
- What to do when there is nonequivalence

Measurement in multigroup modelling

So far we have assumed **measurement equivalence** across groups, i.e.

$$p(\mathbf{y}|\boldsymbol{\eta}, \mathbf{x}) = p(\mathbf{y}|\boldsymbol{\eta})$$

or in multiple-group notation

$$p^{(g)}(\mathbf{y}|\boldsymbol{\eta}) = p(\mathbf{y}|\boldsymbol{\eta}).$$

In this session we relax this assumption, to allow for the possibility of **nonequivalence** of measurement.

We focus on models with one latent variable η. Measurement issues in models with multiple latent variables are similar.

We use mostly two simple examples, each with two groups only.

Latent class example: Attitudes to immigration

Data from Round 5 of European Social Survey (2010).

"To what extent do you think [country] should allow people [type] to come and live here?"

with three questions on different [types] of people:

- **same**: *...of the same race or ethnic group as most [country]'s people..."*
- **differ**: *...of a different race or ethnic group from most [country] people..."*
- **poor**: *...from the poorer countries outside Europe..."*

Response options for each: Allow (1) *Many (paljon, många)*, (2) *Some (melko paljon, en del)*, (3) *A few (vähän, några få)*, (4) *None (ei lainkaan, inte tillåta några)*.

Compare data for two countries, **Finland** and **Sweden**.

Latent trait example: Attitudes to abortion

From the British Social Attitudes Survey: *"Here are a number of circumstances in which a woman might consider an abortion. Please say whether or not you think the law should allow an abortion in each case."* (1=Yes, 2=No) :

1. The woman decides on her own that she does not wish to have the child. [WomanDecide]
2. The couple agree that they do not wish to have the child. [CoupleDecide]
3. The woman is not married and does not wish to marry the man. [NotMarried]
4. The couple cannot afford any more children. [CannotAfford]

Compare data from two different *years* of the survey, **1986** and **2004**.

Measurement equivalence and nonequivalence

If measurement equivalence does not hold for an item y_j, the response probabilities

$$\pi_{jl}^{(g)}(\eta) = P^{(g)}(y_j = l|\eta)$$

depend on group g.

- In other words, two subjects who have the same value of η still have different response probabilities if they belong to different groups.

If nonequivalence is present but ignored, it can in principle distort conclusions about the structural model.

- I.e. part of observed differences in response distributions will be due to measurement differences rather than differences in latent distributions.

This is usually (but not always) the main reason to care about nonequivalence of measurement, i.e. as a nuisance to be managed rather than of interest in itself.

Causes of nonequivalence

- Nonequivalence implies that respondents from different groups react differently to an item.
- This may be because the wording is interpreted differently, because of translation.
  - e.g. the qualifiers of the response options in Finnish and Swedish above
- But it may also be because the content of the item genuinely has a different meaning/relevance to different respondents.
  - e.g. in a crime survey: "How worried are you about being attacked due to your ethnic origin?" This functions differently for ethnic minority and majority respondents.
  - This kind of nonequivalence is not limited to cross-national surveys.
- Various combinations and degrees of these are easily imaginable.

Dealing with nonequivalence

We will briefly discuss the following topics:

- Definition or measurement nonequivalence and different types of it.
- How to detect it
- What to do about it.

Construct equivalence

- An item y_j is said to possess **construct equivalence** across groups if it measures the same consruct in every group.
- This is not a formal definition in terms of statistical models.
- With models, construct equivalence can be (somewhat informally) assessed as follows:

  1. Fit models for a set of items separately in each group.
  2. If the same model fits well in each group, the items possess construct equivalence as measures of the latent variables in the model.

  Here "same" means that the number of latent variables (and latent classes) is the same, and measurement models suggest the same interpretation for the latent variables.

Construct equivalence: Immigration example

Separate 4-class models, response probabilities for Finland/Sweden:

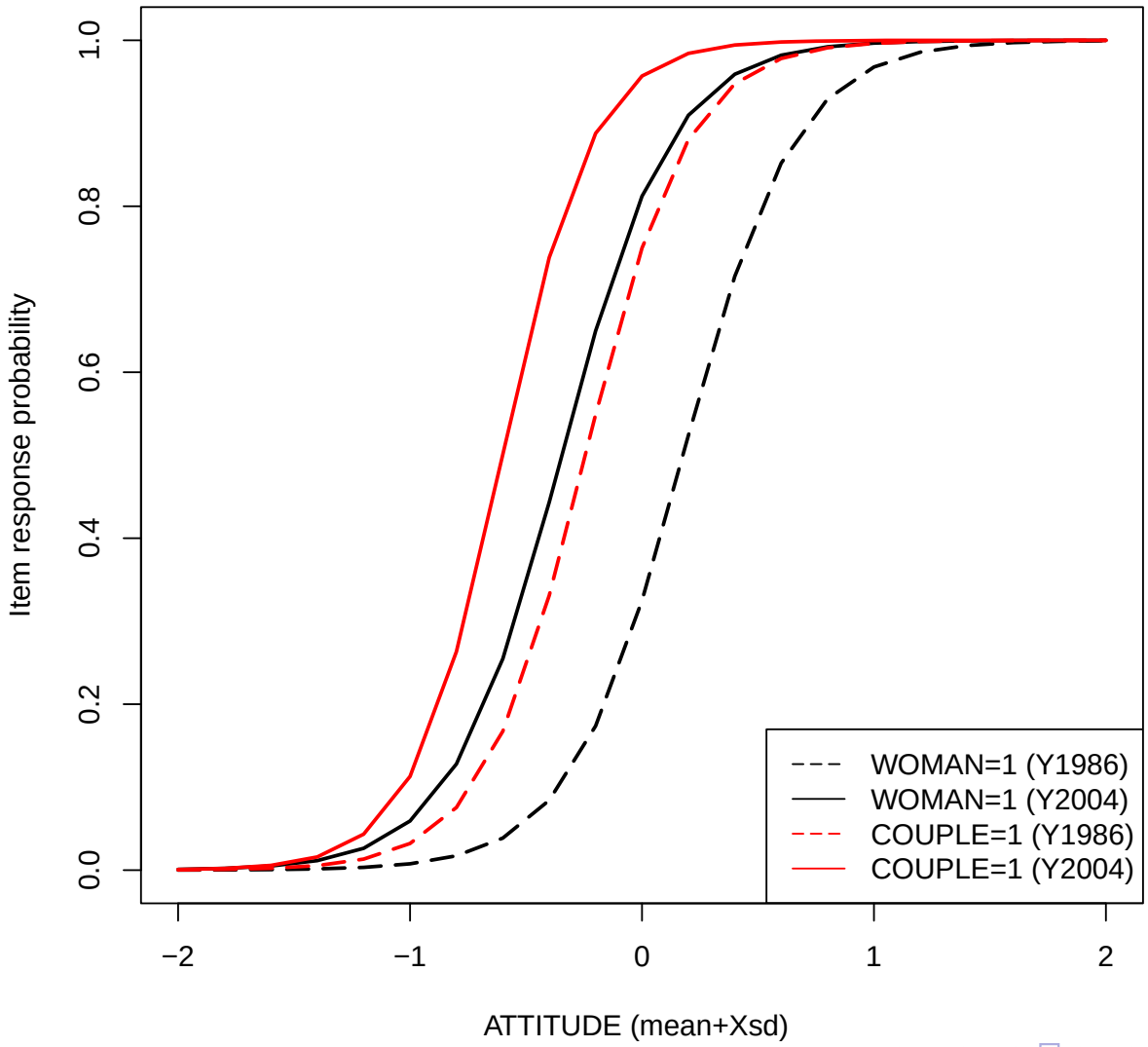
		Latent class			
Item	Level	'Many'	'Some'	'A few'	'None'
same	many	.93/.99	.10/.06	.04/.09	.05/.10
	some	.04/.01	.87/.92	.28/.31	.04/.11
	a few	.03/.00	.03/.01	.68/.60	.51/.20
	none	.01/.00	.00/.00	.00/.00	.40/.59
differ	many	.97/.99	.01/.01	.00/.00	.00/.00
	some	.00/.01	.92/.98	.01/.01	.02/.12
	a few	.02/.00	.07/.01	.96/.98	.01/.00
	none	.00/.00	.01/.00	.03/.00	.97/.88
poor	many	.70/.91	.01/.04	.00/.00	.00/.00
	some	.20/.09	.65/.89	.01/.15	.01/.10
	a few	.09/.00	.33/.06	.83/.76	.08/.07
	none	.01/.00	.01/.00	.16/.09	.91/.82

Construct equivalence: We can interpret the classes similarly in both countries.

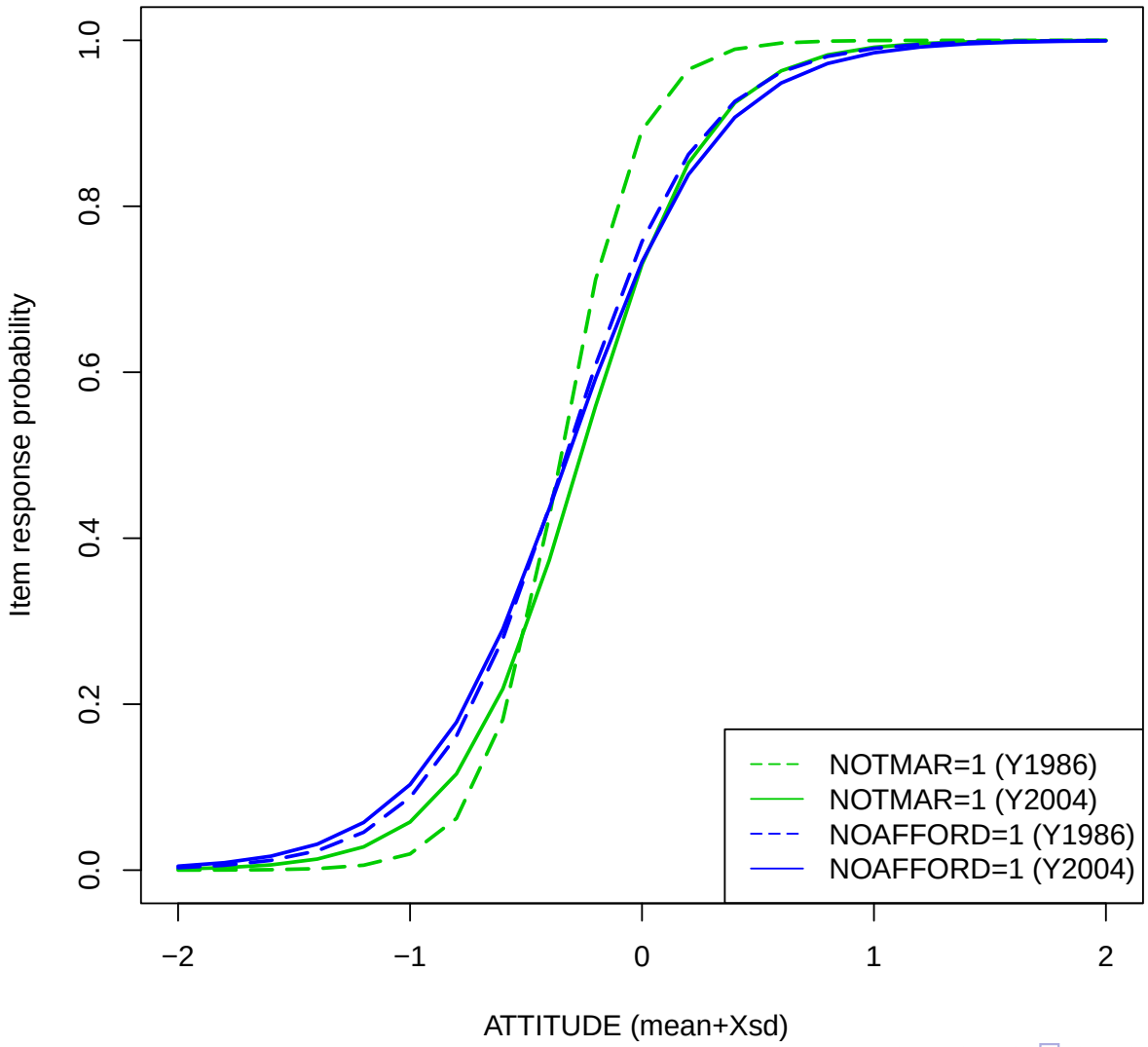
Construct equivalence: Abortion example

- On the next slides, ICC curves for the response "Yes" from 1-trait models fitted separately to the 1986 and 2004 data.
- Both models fit well individually.
- In both, the trait can clearly be interpreted as support for legalised abortion. So construct validity seems satisfied.
- This is really all we can say based on these plots. This is because in the two models the trait is scaled to have mean 0 and variance 1 in each of the groups separately. So the latent scales and thus the shapes of the item response curves are not directly comparable.

Construct equivalence: Abortion example



Construct equivalence: Abortion example



Measurement equivalence within a model

If construct equivalence is judged to hold, we can then formally fit multigroup models for the items, as measures of the same latent variables in each group.

Measurement equivalence is then defined in terms of the measurement models of this model:

- Measurement equivalence holds if $p^{(g)}(y_j|\boldsymbol{\eta}) = p(y_j|\boldsymbol{\eta})$ for all items j.
- There is nonequivalence if $p^{(g)}(y_j|\boldsymbol{\eta})$ varies with g for some j. Sometimes non-equivalence is said to be "partial" if it affects some but not all of the items.

There are further variants depending on which parameters of a measurement model vary.

Level of measurement equivalence can be assessed by comparing fitted models with different constraints, using likelihood ratio tests and other methods of model assessment.

Identifiability of non-equivalence models

For latent trait models, a multigroup model is not identifiable if all items have nonequivalence.

- Then models where the distributions of the traits are the same and not the same across groups are indistinguishable, so we cannot answer comparative questions about the structural model.

To avoid this, identifiability constraints need to be imposed:

- Enough constraints on some parameters in one group to identify the latent scale(s) in it, as in the single-group case (e.g. $(\kappa^{(1)}, \phi^{(1)}) = (0, 1)$ in the one-trait model).
- One item per trait must have equivalence, i.e. in the one-trait model $p^{(g)}(y_j|\eta) = p(y_j|\eta)$ for one j.

The equivalent items serve as "anchors" which establish a common scale for η across groups.

Identifiability of non-equivalence models

For latent class models, no constraints on the structural model and no anchor items are needed for identifiability.

- Apart from the number of classes, and subject to arbitrary renumbering of the classes.
- In other words, the separate modelling used for assessing construct equivalence is also multigroup modelling with complete nonequivalence.
- We can still claim that the latent variable is the same across groups if the measurement patterns are similar enough. The probabilities of the classes can then be compared across groups.

But this is not very comfortable for interpretation, nor is latent trait modelling with just one equivalent item. In practice we would prefer models with partial nonequivalence in at most a few of the items.

Measurement models in multigroup models

For example, consider a model for a binary y_j given latent trait η:

$$\pi_{j1}^{(g)}(\eta) = \frac{\exp(\tau_j^{(g)} + \lambda_j^{(g)}\eta)}{1 + \exp(\tau_j^{(g)} + \lambda_j^{(g)}\eta)}.$$

This has two parameters, **intercept** $\tau_j^{(g)}$ and **loading** $\lambda_j^{(g)}$.

All of the parameters of other measurement models (multinomial and ordinal for non-binary items, in models with multiple traits, and in latent class models) are also either loadings or intercepts, so we can illustrate the concept in general with the binary model.

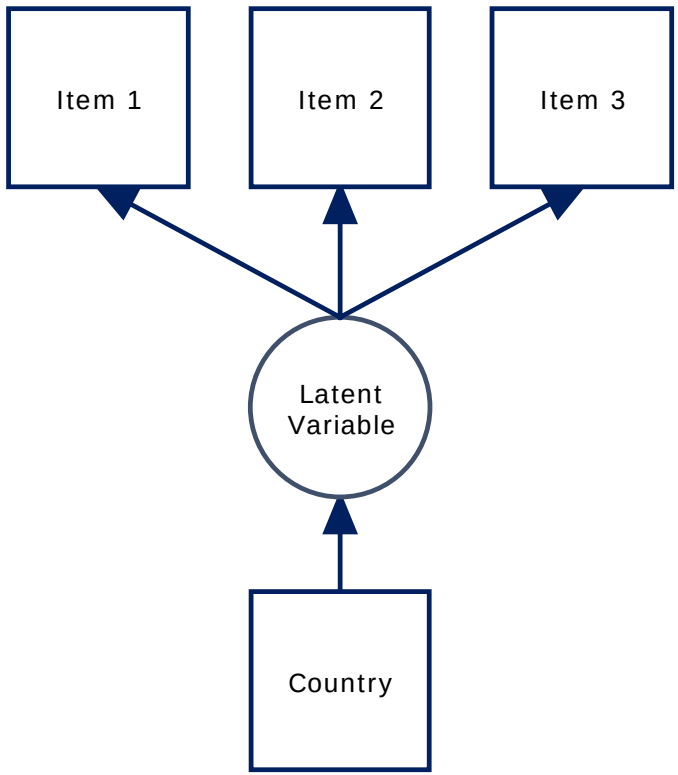
Which parameters are equivalent?

For a single item, we consider three cases:

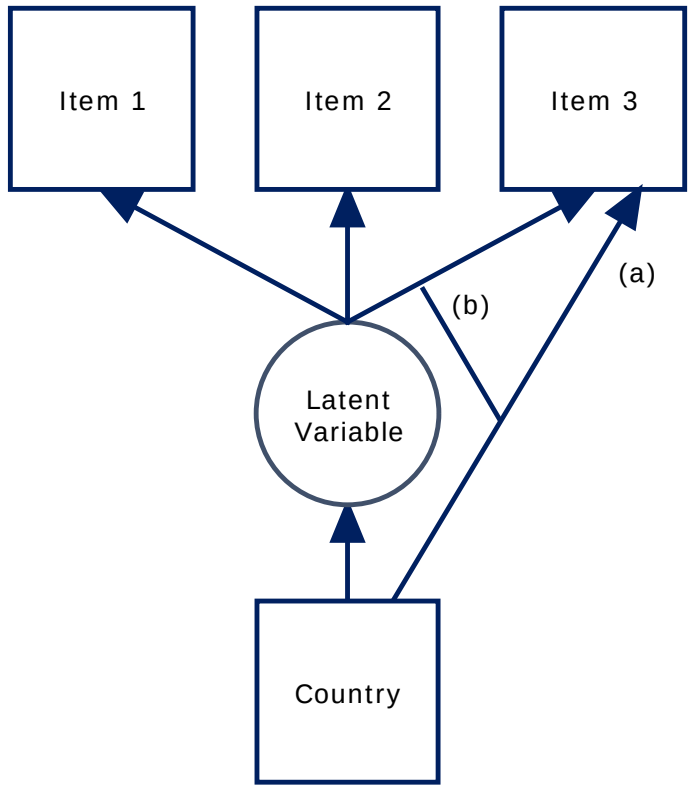
- Equivalence of all parameters ($\tau_j^{(g)} = \tau_j$ and $\lambda_j^{(g)} = \lambda_j$) — see (1) on the next slide.
- Non-equivalence in intercepts but equivalence in loadings ("direct effect", $\lambda_j^{(g)} = \lambda_j$) — see (2) with (a) only.
- Non-equivalence in both intercepts and loadings ("interaction") — (2) with (a) and (b).

Which parameters are equivalent?

(1)



(2)



Which parameters are equivalent?

- The interaction model of measurement states that the association between latent variable and an item varies between groups
  - In a trait model, the "discrimination" of the item varies.
  - In a class model, magnitude and nature of the non-equivalence may be different in different latent classes.
- The direct-effect measurement model states that the between-group differences in response probabilities are the same (on a logit scale) at all values of the latent trait or class.
  - The "difficulty" of an item varies, but discrimination does not.
- For latent class models we may also consider a variant of the interaction model where response probabilities are equivalent in some latent classes but not in others.

Assessing measurement equivalence

In addition to the choices about structural and measurement models discussed previously, we now have a range of possible models with different levels of measurement equivalence.

In principle we can proceed as before: Use model selection statistics to try to identify well-fitting models for the data.

- Tests of measurement equivalence involve nested models which can be compared with likelihood ratio tests (e.g. $\lambda_j^{(1)} = \cdots = \lambda_j^{(G)}$ vs. not).
- For bivariate marginal residuals we now have them both overall (calculated from two-way-tables summed over group too) and given group (calculated from two-way-tables within each group).
  - Sometimes we may see that the overall residuals are small but group-specific ones large. This is strong evidence that the measurement models are inadequate (too equivalent).

Example: Attitudes to immigration

Part of Mplus input for Model 7 on the next page:

```
Model:
    %overall%
        class on gr19;
        poor ON gr19;
    %class#2%
!        poor ON gr19; ! Uncomment for Model 6
    %class#3%
!        poor ON gr19; ! Uncomment for Model 6
    %class#4%
!        poor ON gr19; ! Uncomment for Model 6
```

Example: Attitudes to immigration

Model assessment statistics for various models:

				%Resids>4 (out of 64)			#large sums for item*item (out of 3)		
	Model	**AIC**	**BIC**	**All**	**FIN**	**SWE**	**All**	**FIN**	**SWE**
1	5-cl equivalent	14522	14846	2	6	10	0	1	1
	4-cl models:								
2	separate	14457	14934	0	0	0	0	0	0
3	equivalent	14684	14941	0	15	17	0	2	2
	noneq. in:								
4	*same*	14672	15002	0	15	12	0	2	2
5	*differ*	14689	15020	0	17	17	0	2	2
6	*poor*	14478	14808	0	6	4	0	0	0
7	*poor*, direct effect only	14470	14745	0	10	2	0	0	0

LR test of 3 vs. 7 has $p < 0.0005$ and 7 vs. 6 has $p = 0.39$.

We consider Model 7, which has nonequivalence in the interept term of measurement model of item **poor**.

Example: Attitudes to immigration

		Probability of latent class:			
		'Many'	'Some'	'A few'	'None'
	Finland	.06	.32	.54	.08
	Sweden	.35	.54	.09	.01
Item	Level	Item response probabilities (Fin/Swe):			
same	many	.98	.08	.04	.05
	some	.02	.91	.29	.05
	a few	.00	.01	.66	.47
	none	.00	.00	.00	.42
differ	many	.99	.01	.00	.00
	some	.01	.96	.01	.02
	a few	.00	.03	.96	.01
	none	.00	.00	.03	.97
poor	many	.73/.90	.01/.04	.00/.02	.00/.00
	some	.19/.09	.66/.89	.02/.12	.01/.11
	a few	.07/.01	.32/.06	.83/.75	.08/.08
	none	.01/.00	.01/.00	.15/.11	.91/.80

For item **poor**, Finnish response probabilities are shifted in the negative direction compared to Sweden, uniformly across the classes.

Example: Attitudes to abortion

Part of Mplus input for Model 8 on the next page:

```
Variable:
    Classes = year (2);
    Knownclass= year (yearx=1 yearx=2);
Model:
    %overall%
        attitude BY woman* couple notmar noafford;
        [attitude@0]; attitude@1;
    %year#2%
        [attitude];
!        attitude BY notmar noafford; ! Uncomment for Model 7
        [notmar$1]; [noafford$1];
```

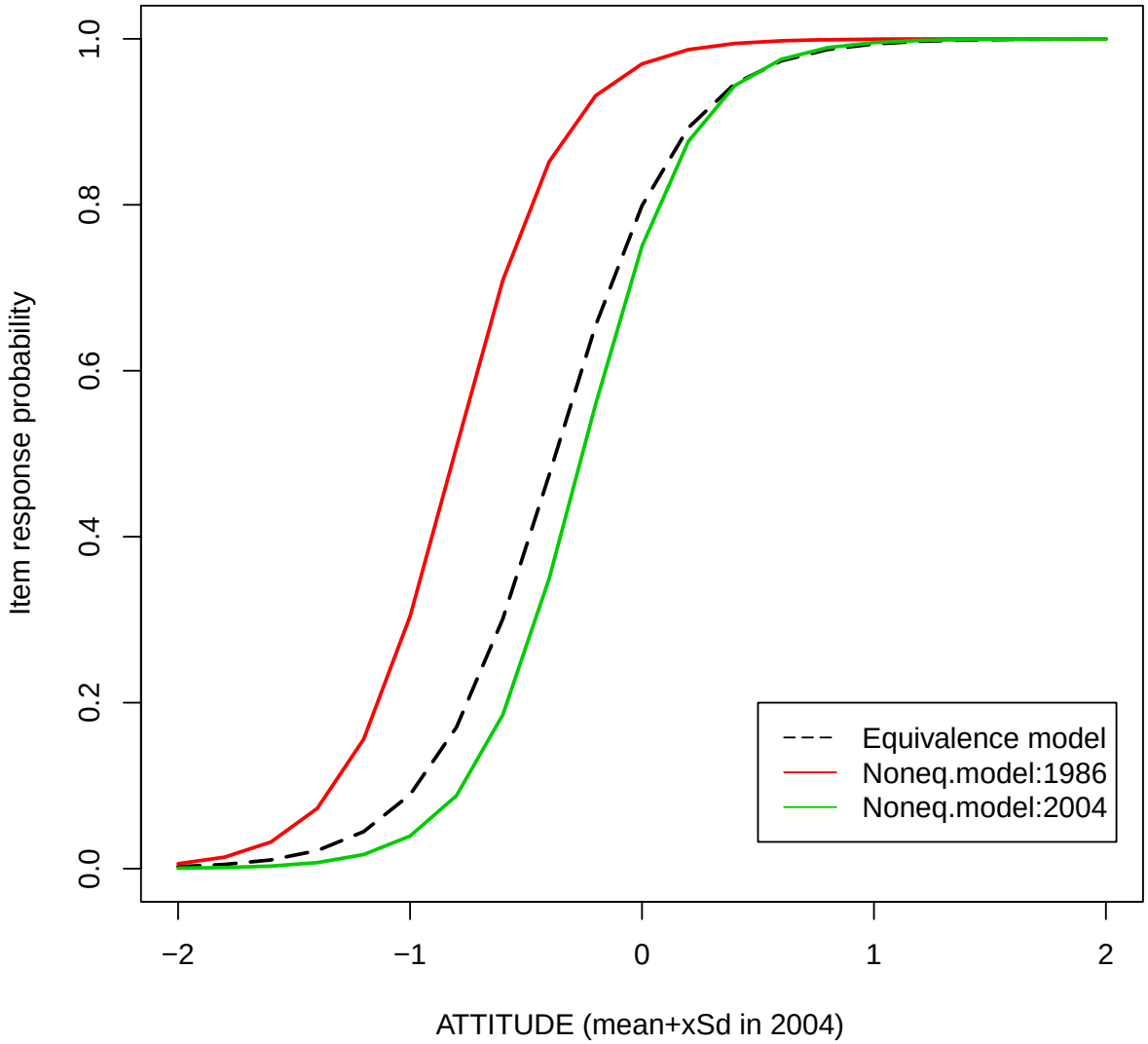
Example: Attitudes to abortion

Model assessment statistics for various ordinal 1-trait models:

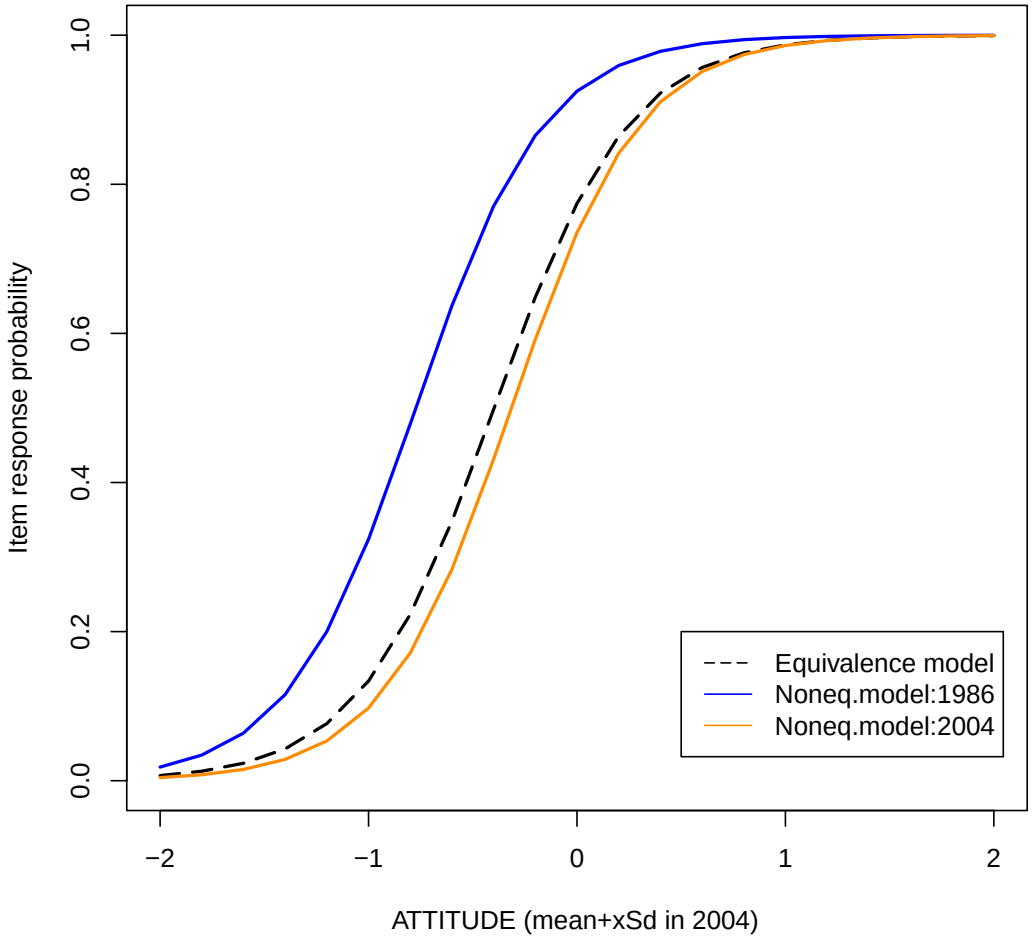
				%Resids>4 (out of 16)			#large sums for item*item (out of 6)		
	Model	**AIC**	**BIC**	**All**	**1986**	**2004**	**All**	**1986**	**2004**
	Equivalence:								
1	Variances different	4466	4518	0	42	29	1	6	5
2	Variances equal	4464	4510	0	42	29	0	6	5
	Nonequivalence in:								
3	*WomanDecide*	4418	4418	4	29	17	1	4	2
4	*CoupleDecide*	4454	4509	4	38	25	0	5	3
5	*NotMarried*	4413	4469	4	29	21	1	3	3
6	*CannotAfford*	4448	4504	4	29	21	1	4	3
7	*NotM & CannotA*	4375	4441	0	0	0	0	0	0
8	*NotM & CannotA*, intercepts only	4376	4432	0	0	0	0	0	0

LR test of 7 vs. 8 has $p = 0.11$. We consider Model 7, which has nonequivalence in the intercept terms of measurement models of items **NotMarried** and **CannotAfford**. In this model, average attitude is 0.45 units more positive (s.e. = 0.08) in 2004 than in 1986, and variance 1 in both years.

Attitudes to abortion: Probability of 'Yes'



Attitudes to abortion: Probability of 'Yes'



In effect, the two items have become "more difficult" from 1986 to 2004.

What to do about nonequivalence?

If we can conclude full measurement equivalence, we are usually happy.

What if not?

1. Omit items and/or groups until equivalence holds for the rest. This is obviously not very attractive.
2. Use a non-equivalence model. This is valid, as long as we believe in the model, i.e. that the same construct is still being measured, even if somewhat differently in different groups. Assuming this is more comfortable if only a few items are nonequivalent.
3. Use the equivalence model anyway. This is convenient, because fitting models with nonequivalence is harder. But conclusions about the structural model may be misleading since the measurement model is poorly specified.

Below we discuss some points related to 2. and 3.

Latent variable scores from multigroup models

From a multigroup model, a prediction of latent variable η given response pattern $\mathbf{y}$ (i.e. trait score or assignment to a latent class) is based on the conditional ('posterior') distribution

$$p^{(g)}(\eta|\mathbf{y}) = \frac{p^{(g)}(\mathbf{y}|\eta)p^{(g)}(\eta)}{\int p^{(g)}(\mathbf{y}|\eta)p^{(g)}(\eta)\,d\eta}$$

This depends on group g in two ways. First, through dependence on the structural model $p^{(g)}(\eta)$. Here we might prefer to replace this with a common distribution, say $p^*(\eta) = \sum_{g=1}^{G} q_g p^{(g)}(\eta)$ where q_g is the proportion of group g in the data.

Latent variable scores from multigroup models

After this change, the posterior distribution for the latent variable is

$$p^{(g)}(\eta|\mathbf{y}) = \frac{p^{(g)}(\mathbf{y}|\eta)p^*(\eta)}{\int p^{(g)}(\mathbf{y}|\eta)p^*(\eta)\,d\eta}$$

This still depends on group g if $p^{(g)}(\mathbf{y}|\eta)$ does,. This is an inevitable implication of nonequivalence:

- If there is measurement nonequivalence, individuals with exactly the same observed responses $\mathbf{y}$ but from different groups will be assigned (predicted) a different value of the latent variable(s).

This is how it should be, if we take the model seriously. But we must be comfortable with this implication if we want to use such a model.

How about those crossnational surveys?

We have shown examples with 2 groups, to introduce basic ideas of measurement equivalence.

However, in many interesting problems the number of groups is much larger, e.g. 20–40 in European cross-national surveys.

We can still do the kinds of analyses discussed above. However, they can become harder and less productive:

- Computations get slower.
- Often model selection statistics suggest that there is clear nonequivalence in most or all items, so that no simple well-fitting models can be found.

More groups: An example

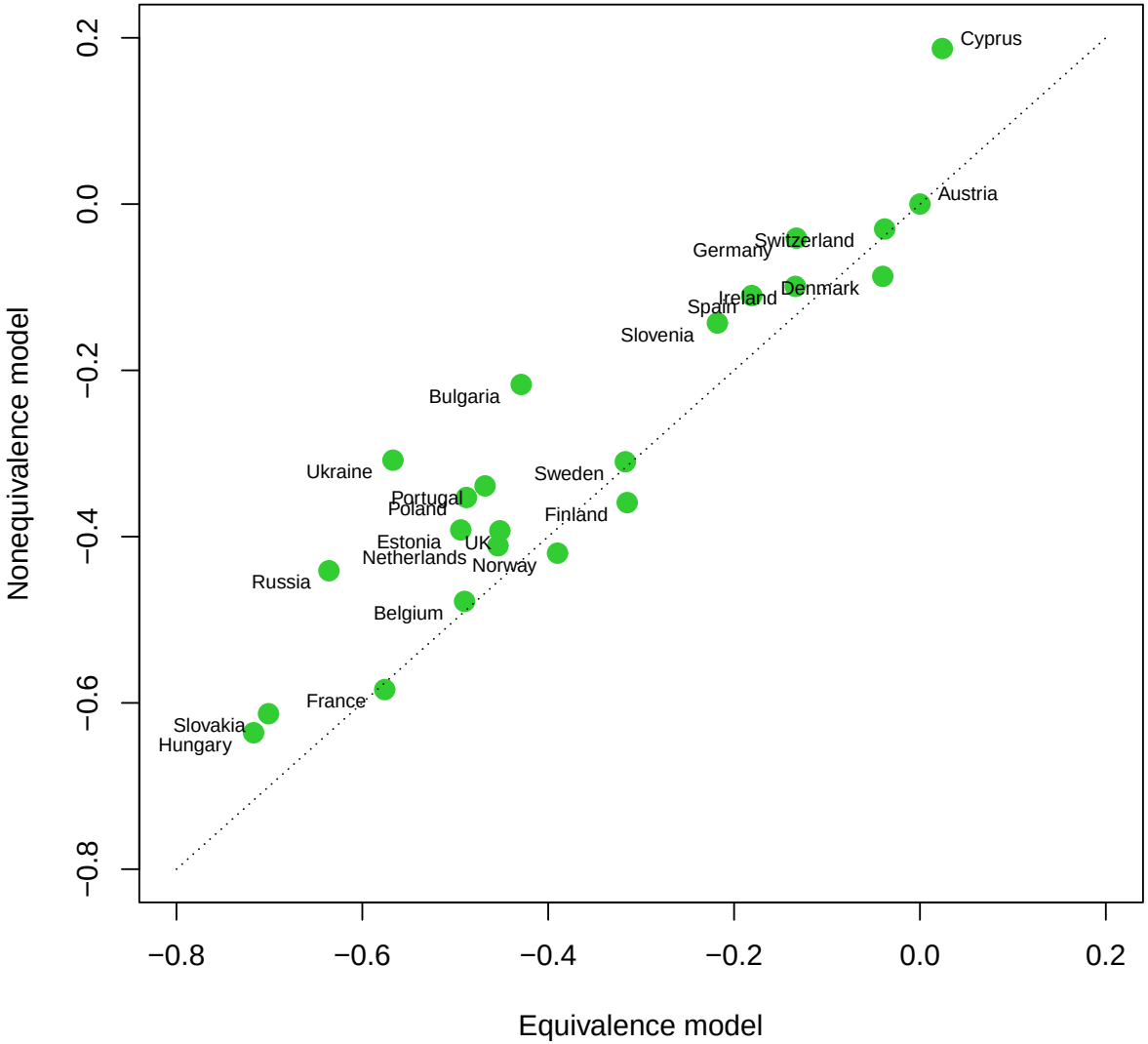
- Happiness example from ESS3 introduced earlier: 4 ordinal items with 5 levels each, 23 countries.
- The 1-trait equivalence model shown earlier has poor fit according to residual statistics: 88% of the overall bivariate residuals, and 37–70% given country, are greater than 4, and sums of the residuals are (very) large for all of the 6 pairs of the variables, overall and for every country.
- Allowing nonequivalence in individual items does not help much. For example, freeing intercepts and loadings in the item *ideal* improves the fit very significantly, but the percentages above remain 83% and 25–65%.
- It may not always be possible to find a model which is both usefully simple and well-fitting according to the statistics.

Sensitivity of conclusions

- What happens if we do not get the measurement model quite right? In particular, what if we simply fit the equivalence model, even when it clearly fits poorly?
- In particular, what happens to conclusions about the structural model, i.e. comparisons of the groups in terms of the latent variables?
- Example: Estimated latent means in the happiness example, from the equivalence model and from a model with one nonequivalent item.
  - Here the broad patterns are unchanged, but there are some changes in the order. Difference between countries are generally smaller in the nonequivalence model.
- It may be that in pratice conclusions are often relatively insensitive to ignoring nonequivalence. However, they will not always be.
- Research into this question of sensitivity is ongoing.

Sensitivity of conclusions: Example

Estimated means of a latent trait, 23 countries:



That is all. Thank you for your attention!

stats.lse.ac.uk/lcat/