Measurement in cross-national comparative research

Tarani Chandola
University of Manchester
tarani.chandola@manchester.ac.uk

Outline

- Systematic and Random measurement error
- Examples of measurement issues in comparative research
- Standard tests of measurement invariance
- Approximate measurement invariance
- An example using ELSA cognition

How to compare measurements in different countries?

Increasing availability of large cross-cultural and cross-country surveys

Increased possibilities to conduct comparative studies.

However, increased the risk of drawing wrong conclusions because of systematic measurement error

Some SF-36 questions

ENERGY AND EMOTIONS:
These questions are about how you feel and how things have been with you during the last 4 weeks. Fo
question, please give the answer that comes closest to the way you have been feeling.

Did you feel full of pep?
- ◯All of the time
- ◯Most of the time
- ◯A good Bit of the Time
- ◯Some of the time
- ◯A little bit of the time
- ◯None of the Time

Have you been a very nervous person?
- ◯All of the time
- ◯Most of the time
- ◯A good Bit of the Time
- ◯Some of the time
- ◯A little bit of the time
- ◯None of the Time

Have you felt so down in the dumps that nothing could cheer you up?
- ◯All of the time
- ◯Most of the time
- ◯A good Bit of the Time
- ◯Some of the time
- ◯A little bit of the time
- ◯None of the Time

Have you felt calm and peaceful?
- ◯All of the time
- ◯Most of the time
- ◯A good Bit of the Time
- ◯Some of the time
- ◯A little bit of the time
- ◯None of the Time

Did you have a lot of energy?
- ◯All of the time
- ◯Most of the time
- ◯A good Bit of the Time
- ◯Some of the time
- ◯A little bit of the time
- ◯ None of the Time

質問9は9問あります。
過去1ヶ月間に、あなたがどのように感じたかの質問です。それぞれの質問について、
一番よくあてはまる番号を選んで下さい。

質問9の1、元気いっぱいでしたか。
「いつも」の場合は1、
「ほとんどいつも」の場合は2、
「ときどき」の場合は3、
「まれに」の場合は4、
「ぜんぜんない」場合は5。

質問9の2、かなり神経質でしたか。
「いつも」の場合は1、
「ほとんどいつも」の場合は2、
「ときどき」の場合は3、
「まれに」の場合は4、
「ぜんぜんない」場合は5。

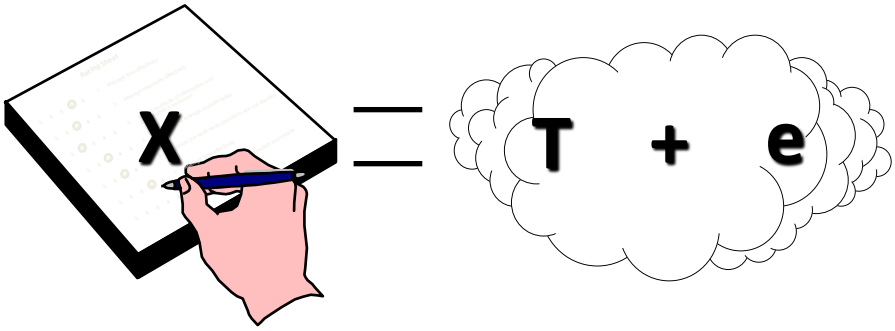
質問9の3、どうにもならないくらい気分が落ち込んでいましたか。
「いつも」の場合は1、
「ほとんどいつも」の場合は2、
「ときどき」の場合は3、
「まれに」の場合は4、
「ぜんぜんない」場合は5。
(回答された数字をご記入下さい)

質問9の4、おちついていておだやかな気分でしたか。
「いつも」の場合は1、
「ほとんどいつも」の場合は2、
「ときどき」の場合は3、
「まれに」の場合は4、
「ぜんぜんない」場合は5。

質問9の5、活力、エネルギーにあふれていましたか。
「いつも」の場合は1、
「ほとんどいつも」の場合は2、
「ときどき」の場合は3、
「まれに」の場合は4、
「ぜんぜんない」場合は5。

True Score Theory

Observed score = True score + Random error

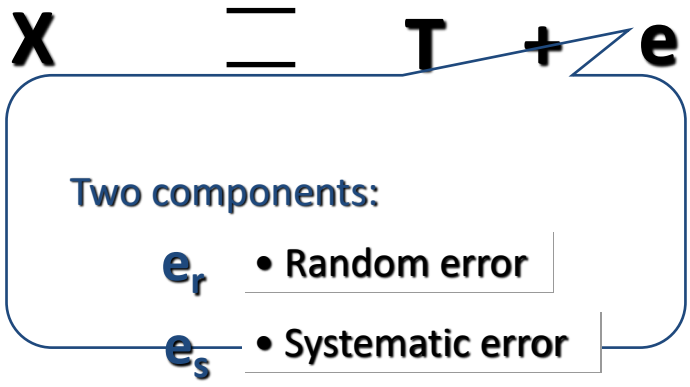


The Error Component

X = T + e

Two components:

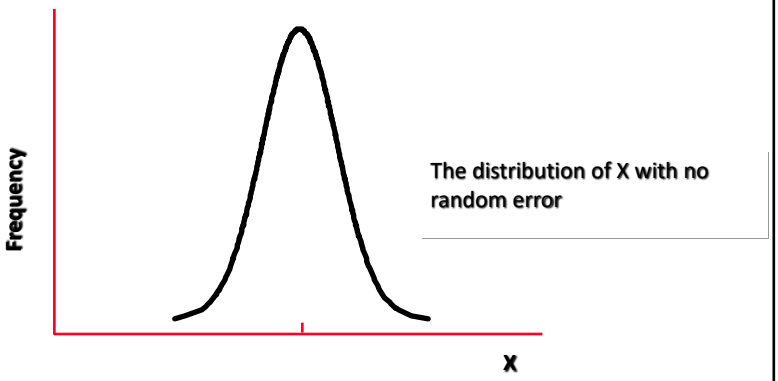
The Error Component



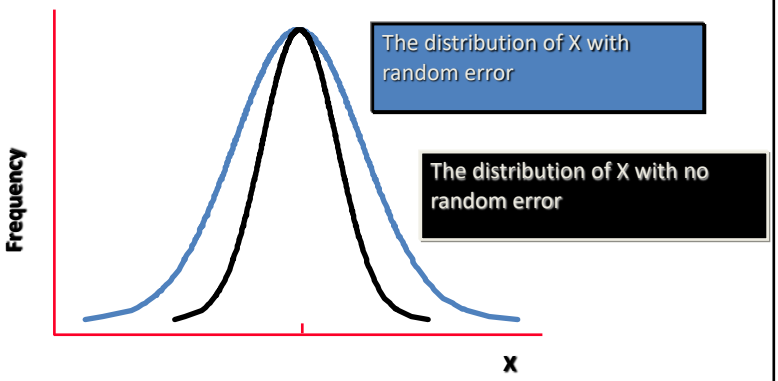
What Is Random Error?

- Any factors that randomly affect measurement of the variable across the sample.
- For instance, each person's mood can inflate or deflate performance on any occasion.
- Random error adds variability to the data but does not affect average performance for the group.

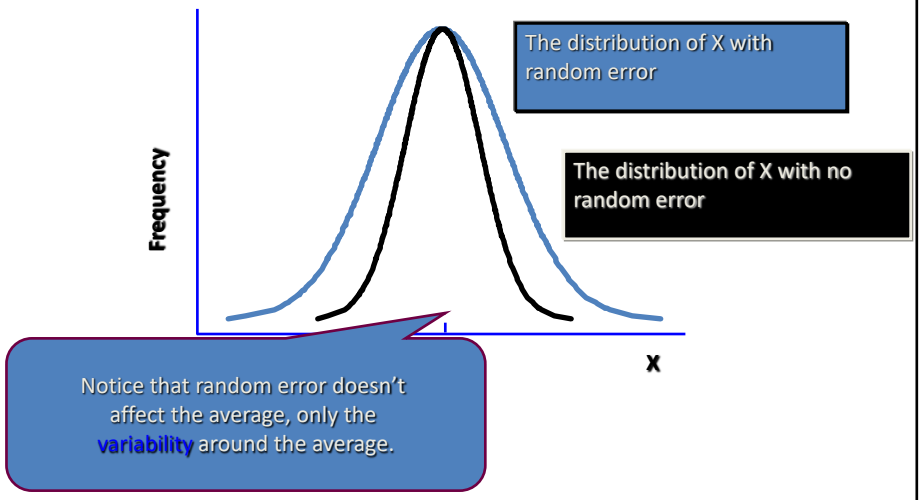
Random Error



Random Error



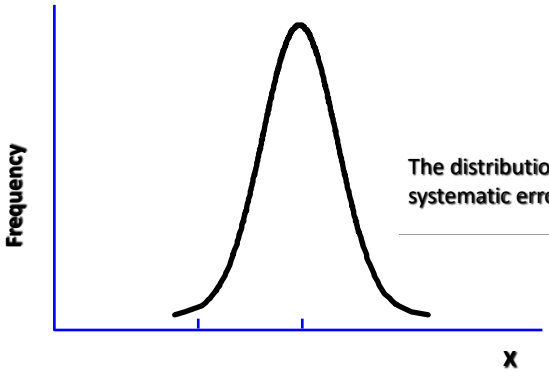
Random Error



Any factors that systematically affect measurement of the variable across the sample.

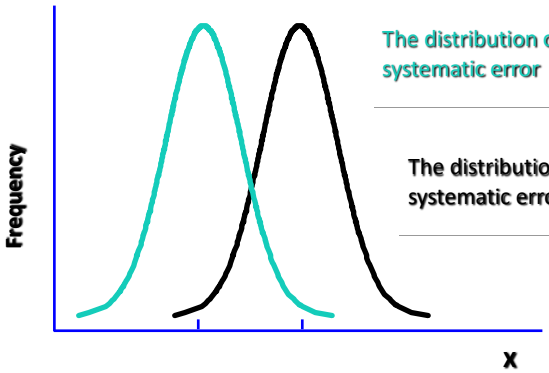
- Systematic error = bias.
- For instance, asking questions that have a cultural or normative bias
- Systematic error does affect average performance for the group.

Systematic Error



The distribution of X with no systematic error

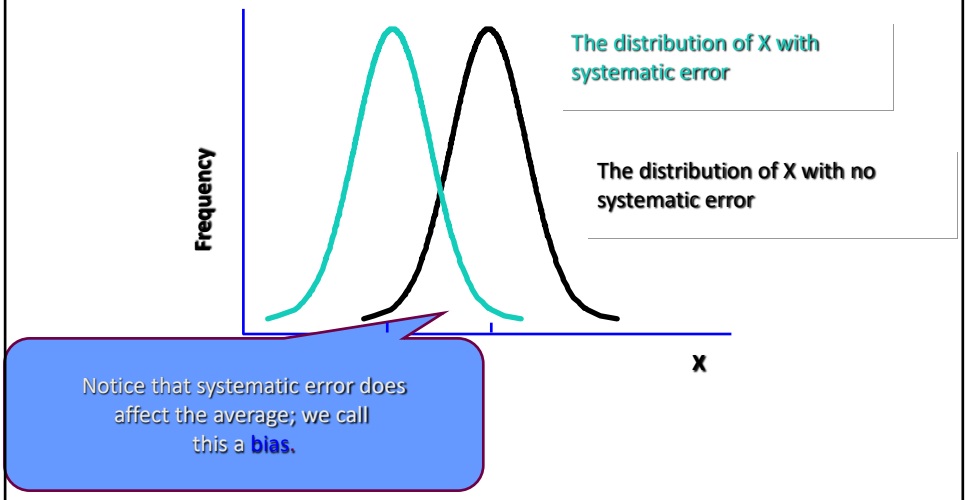
Systematic Error



The distribution of X with systematic error

The distribution of X with no systematic error

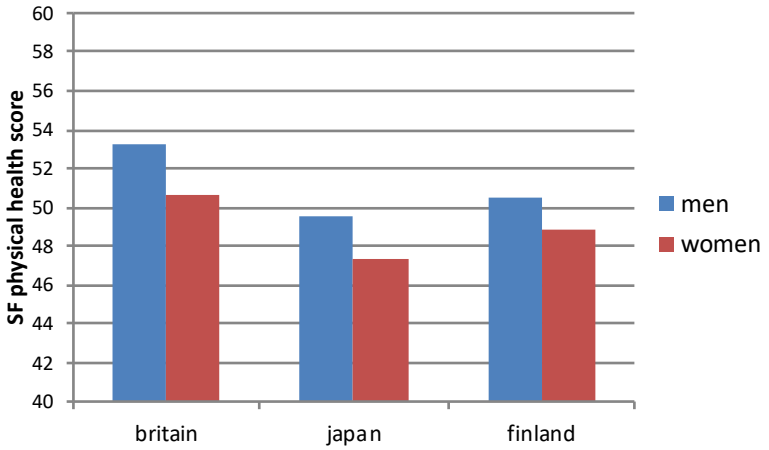
Systematic Error



Outline

- Systematic and Random measurement error
- Examples of measurement issues in comparative research
- Standard tests of measurement invariance
- Approximate measurement invariance
- An example using ELSA cognition

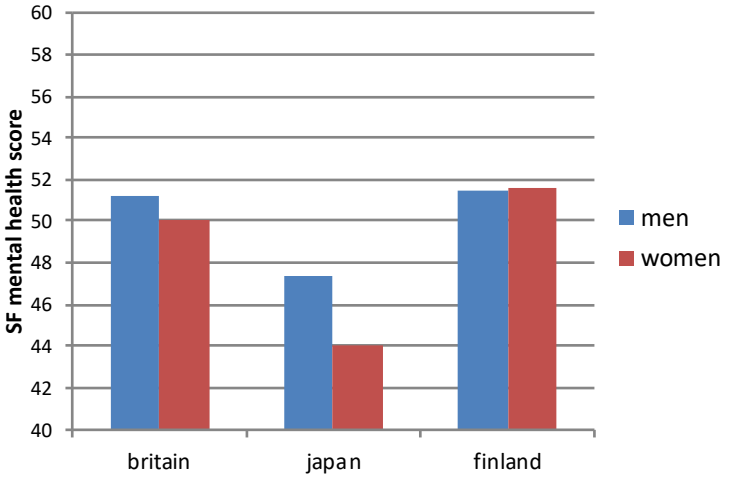
Mean SF physical health score: Britain, Japan and Finland civil servants



What we learn from British, Finnish, and Japanese civil servants study and the role of social democracy in reducing socioeconomic inequalities in health: A response to Bosma

Michikazu Sekine [a,*], Tarani Chandola [b], Pekka Martikainen [c], Michael Marmot [b], Sadanobu Kagamimori [a]

Mean SF mental health score: Britain, Japan and Finland civil servants



What we learn from British, Finnish, and Japanese civil servants study and the role of social democracy in reducing socioeconomic inequalities in health: A response to Bosma

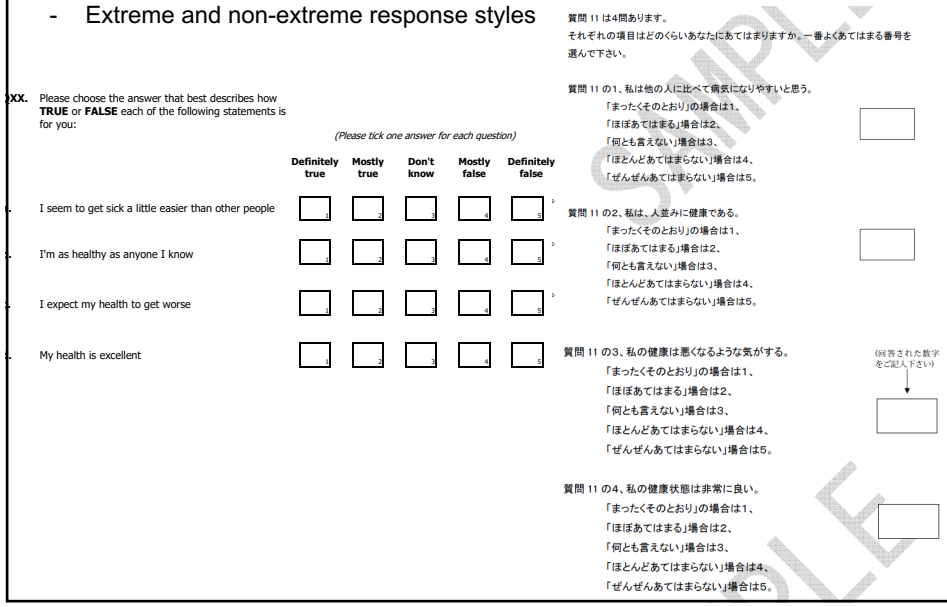
Michikazu Sekine [a,*], Tarani Chandola [b], Pekka Martikainen [c], Michael Marmot [b], Sadanobu Kagamimori [a]

Interpretation problems

- Can the study infer which country has better mental health functioning?
- How to distinguish between systematic and random measurement error?

Measurement problems SF-36

- Cultural norms
- Extreme and non-extreme response styles



XX. Please choose the answer that best describes how **TRUE** or **FALSE** each of the following statements is for you:

(Please tick one answer for each question)

	Definitely true	Mostly true	Don't know	Mostly false	Definitely false
I seem to get sick a little easier than other people	1	2	3	4	5
I'm as healthy as anyone I know	1	2	3	4	5
I expect my health to get worse	1	2	3	4	5
My health is excellent	1	2	3	4	5

質問 11 は4問あります。
それぞれの項目はどのくらいあなたにあてはまりますか。一番よくあてはまる番号を選んで下さい。

質問 11 の1、私は他の人に比べて病気になりやすいと思う。
「まったくそのとおり」の場合は1、
「ほぼあてはまる」場合は2、
「何とも言えない」場合は3、
「ほとんどあてはまらない」場合は4、
「ぜんぜんあてはまらない」場合は5。

質問 11 の2、私は、人並みに健康である。
「まったくそのとおり」の場合は1、
「ほぼあてはまる」場合は2、
「何とも言えない」場合は3、
「ほとんどあてはまらない」場合は4、
「ぜんぜんあてはまらない」場合は5。

質問 11 の3、私の健康は悪くなるような気がする。
「まったくそのとおり」の場合は1、
「ほぼあてはまる」場合は2、
「何とも言えない」場合は3、
「ほとんどあてはまらない」場合は4、
「ぜんぜんあてはまらない」場合は5。

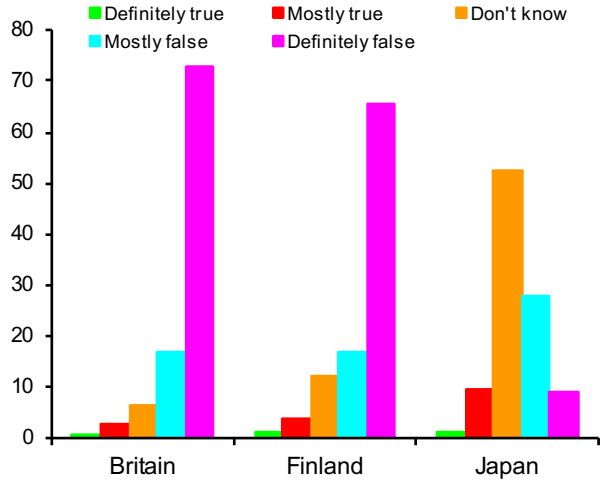
(回答された数字をご記入下さい)

質問 11 の4、私の健康状態は非常に良い。
「まったくそのとおり」の場合は1、
「ほぼあてはまる」場合は2、
「何とも言えない」場合は3、
「ほとんどあてはまらない」場合は4、
「ぜんぜんあてはまらない」場合は5。

Measurement problems SF-36

Percentage distribution of responses to
'I seem to get sick a little easier than other people'

質問 11 の 1、私は他の人に比べて病気になりやすいと思う。
「まったくそのとおり」の場合は1、
「ほぼあてはまる」場合は2、
「何とも言えない」場合は3、
「ほとんどあてはまらない」場合は4、
「ぜんぜんあてはまらない」場合は5。



How to compare measurements in different countries?

When just one measurement item (question), you cannot distinguish between systematic and random error

You can compare how the item differs between countries in terms of predicted differences

Self-rated health

QXX. In general would you say your health is: *(Please tick one)*

Excellent	1	yoi
Very good	2	maa yoi
Good	3	futsu
Fair	4	amari yokunai
Poor	5	yokunai

- All reasonable translations of ***fair*** were indistinguishable from the translations of ***good***. Term used- ***amari yokunai*** (***somewhat not good)***

- ***Poor*** is not the same as ***qarui*** (***bad or terrible***)- not appropriate.
Yokunai (***not good)*** - used instead

Does self rated health measure the same concept across countries? Insights from a comparison of older adults in England and Japan.

Benjamin D Williams[1] MSc, Tarani Chandola[1] PhD, Noriko Cable PhD[2]
1. Cathie Marsh Institute for Social Research (CMIst), University of Manchester
2. Research Department of Epidemiology and Public Health, University College London

Self-Rated Health (SRH) is predictive of morbidity and mortality, correlates well with objective measurements of physical function and is simple to use in multidisciplinary surveys.

However, it may not be comparable between countries which may wish to contrast health policies due to linguistic, cultural or health differences

Does self rated health measure the same concept across countries? Insights from a comparison of older adults in England and Japan.

Methods:

- English Longitudinal Study of Ageing (ELSA; 2004, 2008 and 2012) and the Japanese Study of Aging and Retirement (JSTAR; 2007, 2009 and 2011),
- n=10, 174 ELSA participants and n=4279 JSTAR participants
- SRH was measured on a 5 point Likert scale which was dichotomised into 1-3 being good health and 4-5 bad health.
- Grip strength (in kilograms) was mean centred by gender and country for analysis.
- Centre for Epidemiology Scale of Depression was used for depression and dichotomised into depressed and non-depressed.
- BMI and smoking
- Multilevel binary logistic regression was used to test whether participants' country of residence was associated with odds of fair or poor SRH and whether the country of residence would moderate associations between SRH and grip strength, depression, smoking or BMI.

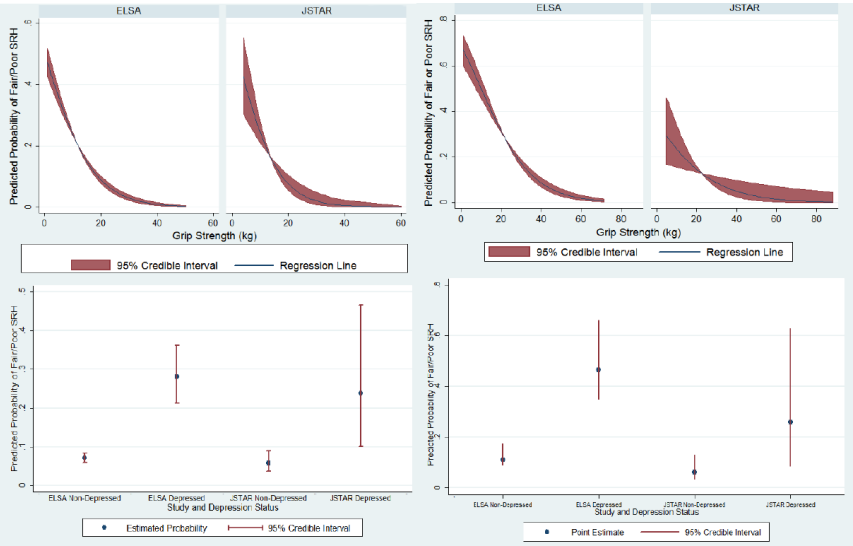
Key Estimates of the Odds of Poor Self Rated Health from fully adjusted Growth Curve Model for each gender

	Women	95% CI		**Men**	95% CI	
	Odds Ratio	Lower	Upper	Odds Ratio	Lower	Upper
Grip Strength (kg)†	0.88	0.86	0.90	0.92	0.91	0.94
Depression (vs non-depressed)	5.49	4.45	6.70	7.11	5.33	9.34
BMI (kg/m^2)†	1.13	1.11	1.15	1.12	1.09	1.15
Smoking Status 1-9 per day	2.74	1.69	4.21	1.16	0.60	2.01
Smoking Status 10 to 19 per day	2.61	1.79	3.68	2.43	1.44	3.85
Smoking Status >= 20 per day	3.76	2.37	5.70	1.97	1.17	3.13
JSTAR (vs ELSA)	0.83	0.60	1.10	0.53	0.36	0.75
JSTAR*Grip Strength	0.98	0.94	1.02	1.02	0.99	1.05
JSTAR*Depression	1.02	0.68	1.48	0.79	0.49	1.22
JSTAR*BMI	0.94	0.90	0.99	0.93	0.87	0.98
JSTAR*Smoking 1-9 per day	0.53	0.16	1.30	1.24	0.34	3.16
JSTAR*Smoking 10-19 per day	0.69	0.31	1.32	0.36	0.16	0.69
JSTAR*Smoking >=20 per day	0.37	0.15	0.78	0.59	0.31	1.05

† Values centred at sample mean, Odds ratio for unit change presented

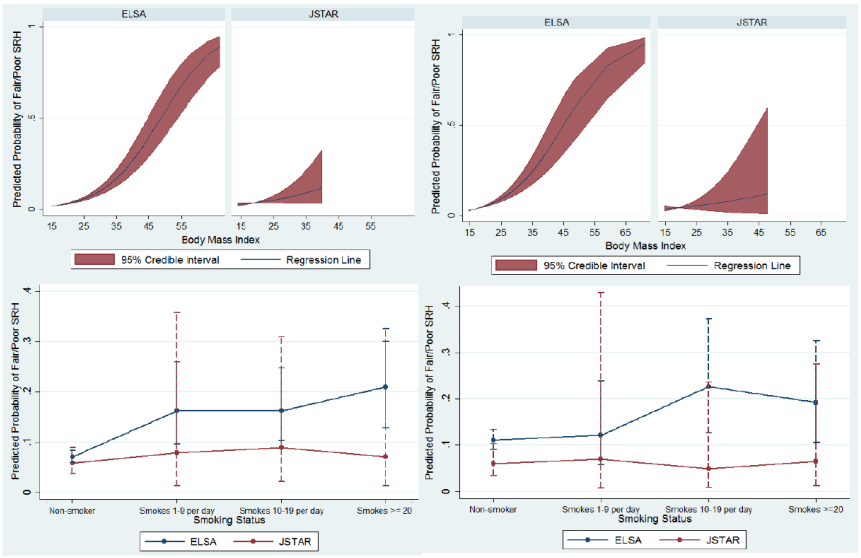
Predicted probability of poor SRH for key covariates by country and gender

Women | Men



Predicted probability of poor SRH for key covariates by country and gender

Women | Men



How to compare measurements in different countries?

Meaning of self-rated health differs between English and Japanese older adults

Cannot directly compare levels of (single item) self-rated health between countries because of systematic bias

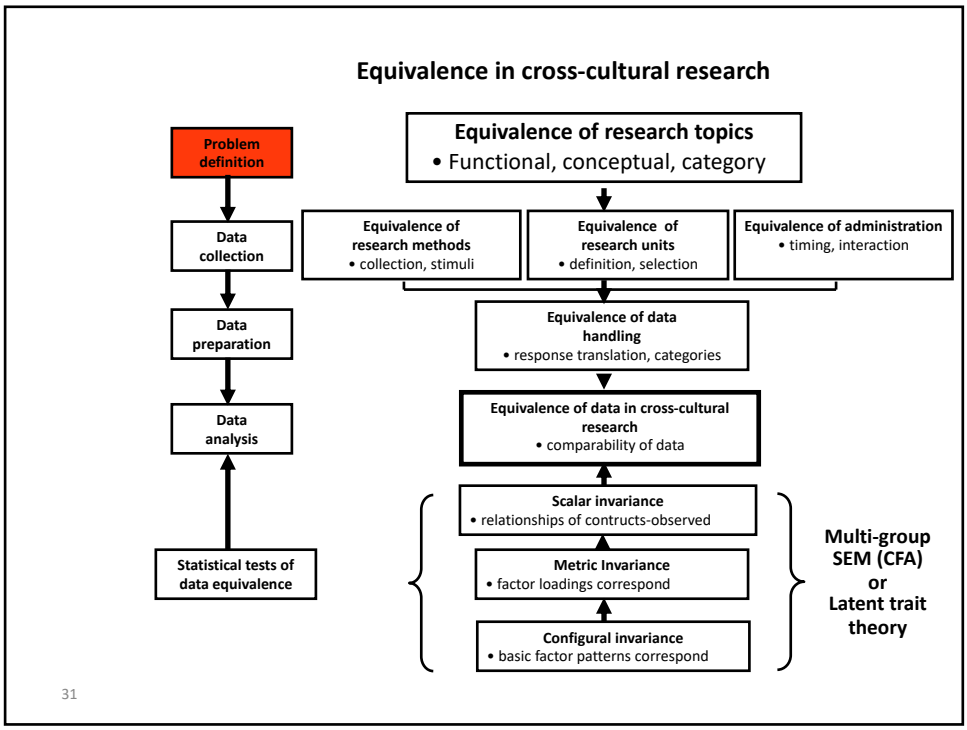
With multi-item (question) scales, possible (in theory) to correct for systematic bias

Test for measurement equivalence of scales (mental health, wellbeing, depression, quality of life) to guarantee that differences across countries are random and not systematic

Unfortunately, a new problem has come up: Many scales do not display high levels of measurement equivalence

Outline

- Systematic and Random measurement error
- Examples of measurement issues in comparative research
- Standard tests of measurement invariance
- Approximate measurement invariance
- An example using ELSA cognition



Statistical tests of data equivalence?

Configural invariance
- ✓ Same items in the latent factors across groups
- ✓ but different factor loadings

Non invariant: cannot compare group means

Metric Invariance
- ✓ Same items in the latent factors across groups
- ✓ Same factor loadings across groups
- ✓ But different intercepts

Weak invariance: cannot compare group means

Scalar Invariance
- ✓ Same items in the latent factors across groups
- ✓ Same factor loadings and intercepts across groups
- ✓ Residuals are different across groups (partial equivalence)

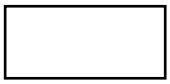
Strong invariance: can compare group means

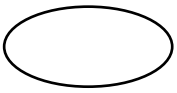
Invariance of latent response
- ✓ Same items in the latent factors across groups
- ✓ Same factor loadings and intercepts across groups
- ✓ Same residuals across groups

32

Terms, Nomenclature, Symbols, and Vocabulary

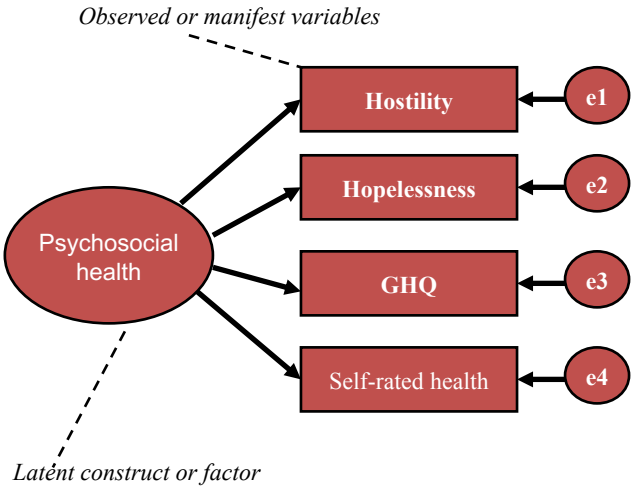
Direct effects 

Observed (or manifest) 

Latent (or factors) 

33

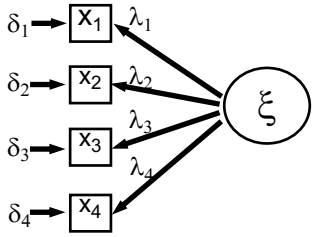
Measurement Model: Confirmatory Factor Analysis



Singh-Manoux, Clark and Marmot. 2002. *Multiple measures of socio-economic position and psychosocial health: proximal and distal measures.*

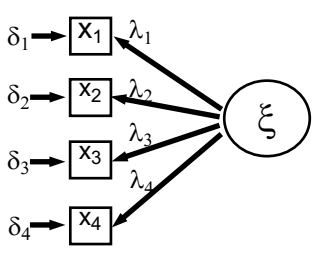
34

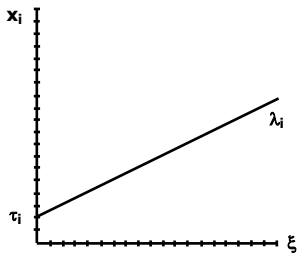
Relationship between latent and observed means



$$x_i = \lambda_i \xi + \delta_i$$

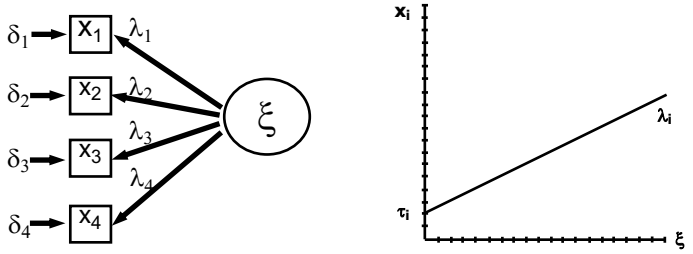
Relationship between latent and observed means





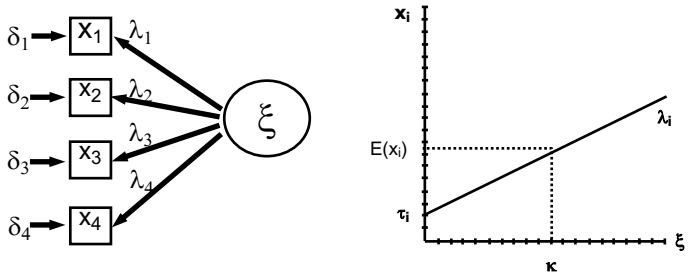
$$x_i = \tau_i + \lambda_i \xi + \delta_i$$

Relationship between latent and observed means



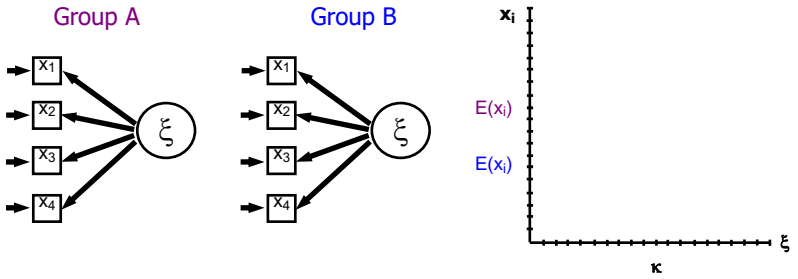
$$E(x_i) = \tau_i + \lambda_i E(\xi) + E(\delta_i)$$

Relationship between latent and observed means

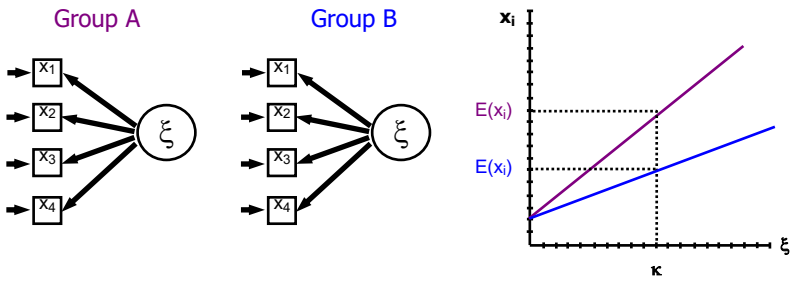


$$E(x_i) = \tau_i + \lambda_i \kappa$$

Group differences in intercepts and factor loadings

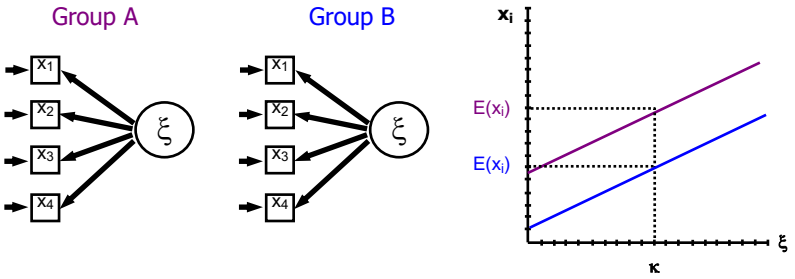


Group differences in intercepts and factor loadings



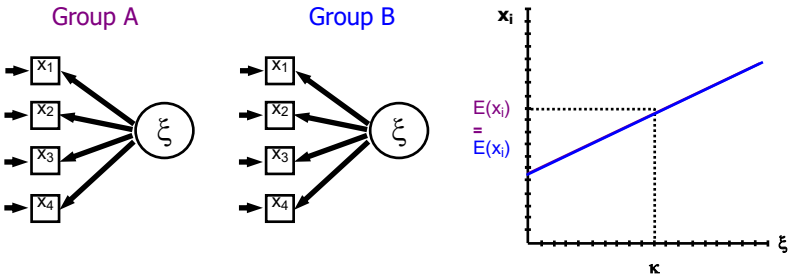
Configurational invariance

Group differences in intercepts and factor loadings



Metric invariance

Group differences in intercepts and factor loadings



Strong invariance

Outline

- Systematic and Random measurement error
- Examples of measurement issues in comparative research
- Standard tests of measurement invariance
- Approximate measurement invariance
- An example using ELSA cognition

Problems with measurement invariance methods

When comparing data from different countries or time points, we want to avoid to paying too much attention to small measurement errors whose effect on substantive conclusions is negligible

Tests for the presence or absence of measurement differences are typically called “measurement invariance tests”, sometimes also known as tests of “differential item functioning”

Techniques to test for measurement invariance are numerous but can be described as broadly falling into two categories: **exact** and **approximate**

Problems with measurement invariance methods

- In the exact methods, the researcher looks for a measurement model in which any "small" measurement differences are assumed to be exactly zero, while "large" differences are left completely free to be estimated from the data (termed 'partial' measurement invariance)
- Methods to establish the fit of such models include chi-square difference testing CFI, RMSEA, and other fit measure comparisons ; and examination of local fit measures such as modification indices (MI)
- However, exact zero constraints are overly strict, especially when there are many groups or time points involved.
- One consequence is a frequent rejection of the exact invariance model, even when the parameter differences are ignorable.
- Another consequence is often a large series of model modifications that appear by chance.
- A ten factor analysis of 21 items over 19 countries yields 380 possible univariate violations of intercept equalities alone. The number of models resulting from all possible combinations of equality restrictions on intercepts and loadings is in the tens of millions.

Approximate measurement invariance

In approximate measurement invariance, small differences in parameters are allowed.

In this "approximate measurement invariance" model, "large" and "small" differences alike are assumed to follow a known distribution of nonzero values. Random effects distributions, multilevel models and strong Bayesian priors have all been used for this purpose.

The idea in all of these techniques is that any smaller differences are automatically accounted for in the model; thus, approximate measurement invariance is primarily designed to deal with the goal of ignoring small differences automatically.

Moreover, the search through all possible combinations of measurement restrictions is replaced by a relatively simple estimation procedure. With many groups and measurement parameters this practical advantage is considerable.

Multigroup confirmatory factor analysis

Given a survey response y_{igj} for respondent *i*, group *g*, and item *j*, a MGCFA measurement model is

$$y_{igj} = \tau_{gj} + \lambda_{gj}\eta_{igj} + \epsilon_{igj},$$

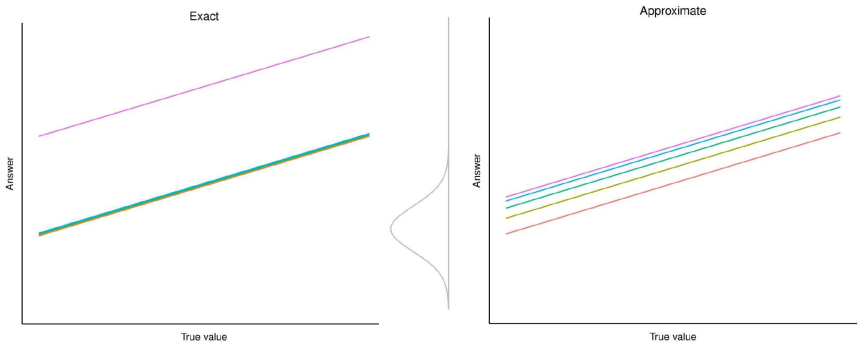
where

- η_{igj} is the unobserved true value (latent variable) for respondent *i;*
- ϵ_{igj} is the unobserved measurement error value (latent variable) for respondent *i;*
- τ_{gj} is the group-specific intercept for item *j;*
- λ_{gj} is the group-specific loading (slope) for item *j.*

Approximate measurement invariance

Kimberley Lek, Utrecht University
Daniel Oberski, Utrecht University
Eldad Davidov, University of Cologne and University of Zurich
Jan Cieciuch, University of Zurich and Cardinal Wyszynski University in Warsaw
Daniel Seddig, University of Zurich
Peter Schmidt, University of Giessen

Response functions (lines) for different groups (colours) under exact/partial (A) vs. approximate (B) measurement invariance models.



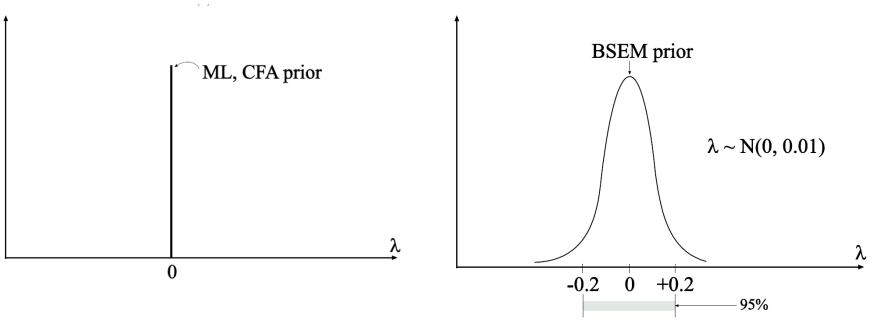
$\tau_{blue,j} = \tau_{green,j} = \tau_{yellow,j} = \tau_{red,j} \neq \tau_{pink,j}$

(A) Exact/Partial measurement invariance

$\tau_{gj} - \tau_{gj'} \sim N(0, \sigma_j)$

(B) Approximate measurement invariance

Maximum Likelihood vs Bayes priors



Bayesian approximate measurement invariance model

- How large the “typical difference” should be?
- Need to appropriately balance the two goals of measurement invariance analysis: accounting for large measurement differences while ignoring the small ones
- A prior variance of ~N(0,0.01) for all differences between loadings, intercepts, and thresholds

Outline

- Systematic and Random measurement error
- Examples of measurement issues in comparative research
- Standard tests of measurement invariance
- Approximate measurement invariance
- An example using ELSA cognition

Example of Bayesian approximate invariance- ELSA cognition

Different dimensions of cognition
Often summed into a single metric score
Sometimes latent variables are used to combine multiple cognitive measures into a single measure

However, cognition changes with age
- Different rates of change in cognitive or physical processes and their associations
- The size of practice effects may also vary between different tests

Any of these may change the strength of the association between the individual cognitive tests and the latent cognitive function over time.

No studies on longitudinal measurement invariance of cognitive tests among older adults

An application of Bayesian measurement invariance to modelling cognition over time in the English Longitudinal Study of Ageing

Benjamin David Williams[1] | Tarani Chandola[1] | Neil Pendleton[2]

Int J Methods Psychiatr Res. 2018;**27**:e1749.
https://doi.org/10.1002/mpr.1749

ELSA cognitive tests

Orientation to time- asking the participant to name the day, year, month and date.

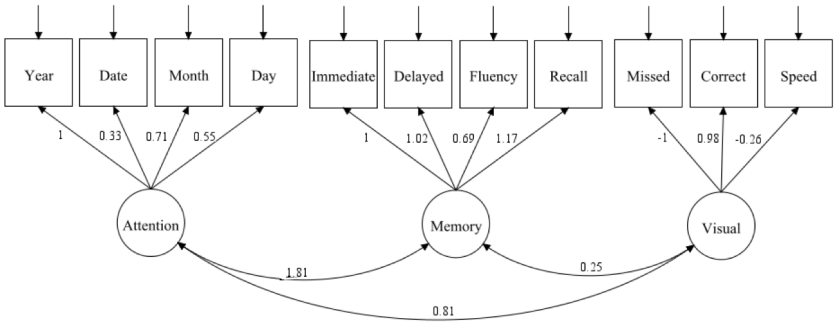
Immediate and delayed verbal recall- a randomly assigned list of 10 common words was played. Delayed recall of the word list was tested after the other cognitive tests were undertaken

The prospective memory task required participants to remember to write their initials in the top corner of a page they were handed.

Semantic fluency was assessed by asking participants to name as many animals as they can in 1 minute.

Letter cancelation task- participants were handed a clipboard with random letters in rows and columns. The aim was to cross out as many of the two target letters as possible in one minute. Participants were asked to complete the task by scanning from left to right as if reading. The number of the last letter reached was used as a measure of processing speed.

CFA Factor Structure for Cognitive Function in the English Longitudinal Study of Ageing.



Mean of correct responses for each cognitive task in ELSA waves 1-5

	Wave				
	1	2	3	4	5
n	11,630	9,066	7,659	6,656	6,535
	Mean				
Immediate	5.4 (1.8)	5.7 (1.8)	5.7 (1.8)	5.7 (1.8)	5.7 (1.9)
Delayed	4.0 (1.8)	4.3 (2.1)	4.4 (2.2)	4.4 (2.2)	4.4 (2.2)
Verbal fluency	19.3 (6.4)	19.8 (6.6)	19.8 (6.8)	20.2 (7.0)	20.2 (7.0)
	Proportion correct (%)				
Year	97.4	98.1	97.5	97.4	97.3
Date	80.6	81.4	80.8	80.8	81.7
Month	97.6	97.7	97.2	97.7	97.8
Day	97.9	97.8	97.6	97.7	97.5
Prospective	79.6	81.3	82.9	84.3	85.8

An application of Bayesian measurement invariance to modelling cognition over time in the English Longitudinal Study of Ageing

Benjamin David Williams[1] | Tarani Chandola[1] | Neil Pendleton[2]

Model fit tests for exact measurement invariance

	χ^2 test versus baseline model	χ^2 test versus less restrictive model
All configural	<0.001	–
Attention strong	<0.001	0.002
Memory weak	<0.001	<0.001
Memory strong	<0.001	<0.001
Both strong	<0.001	<0.001

Factor loadings using Bayesian approximate measurement invariance for both factors at each time point

Item		Approximate invariance factor loadings (0.01 prior variance) Wave 1	Wave 2	Wave 3	Wave 4	Wave 5	Mean
	Year	1	1.021	1.034	1.029	1.045	1.026
Orientation	Date	0.278	0.295	0.298	0.264	0.264	0.28
Factor	Month	0.51	0.54	0.555	0.513	0.516	0.527
	Day	**0.326***	0.387	0.35	0.369	0.377	0.362
	Immediate recall	1	1.013	1.025	1.021	0.985	1.009
Memory	Delayed recall	1.08	1.102	1.101	1.082	1.064	1.086
Factor	Verbal fluency	0.856	0.897	0.896	**0.927***	0.914	0.898
	Prospective mem.	0.88	0.934	0.875	0.911	0.855	0.891
		Approximate invariance intercepts† and thresholds‡ (0.01 prior variance) Wave 1	Wave 2	Wave 3	Wave 4	Wave 5	Mean
	Year‡	−5.887	−5.9	−5.892	−5.898	−5.89	−5.893
Orientation	Date‡	−1.095	−1.099	−1.088	−1.040	−1.062	−1.077
Factor	Month‡	−3.483	−3.446	−3.463	−3.457	−3.476	−3.465
	Day‡	−2.796	−2.853	−2.754	−2.847	−2.811	−2.812
	Immediate recall†	−0.013	**0.053***	**0.033***	−0.013	**−0.055***	0.001
Memory	Delayed recall†	**−0.014***	**0.069***	**0.086***	0.037	−0.013	0.033
Factor	Verbal fluency†	−0.011	**0.009***	−0.013	−0.013	−0.063	−0.018
	Prospective Memory‡	**−0.963***	−1.013	−1.034	−1.064	**−1.086***	−1.032

*Statistically significant using 95% credible interval.

Summary of measurement invariance of ELSA cognitive factors

Exact measurement invariance tests suggests memory and orientation cognitive factors are not comparable across ELSA waves

However, approximate measurement approach identified small but significant non-invariance in the factor loadings of the memory and attention factors

We can assume strong longitudinal measurement invariance in the attention/orientation factor and weak invariance in the memory factor

Overall Summary- 1

- Approximate measurement invariance as a possible solution to the problem of measurement equivalence in cross-national comparative studies
- Instead of restricting the differences between all measurement parameters (i.e., factor loadings, intercepts) to be exactly zero, approximate measurement invariance assumes that these differences follow a (normal) distribution with mean zero and small variance.
- This variance can either be estimated from the data or be fixed in advance by the researcher.
- The latter is known as 'Bayesian' approximate measurement invariance and can be fitted with standard software.

Overall Summary-2

- Approximate measurement invariance seems especially advantageous when
- (1) the number of groups or repeated measurements is large,
- (2) there are many small differences in intercepts and factor loadings and
- (3) differences cancel each other out both within and between groups. Exact measurement invariance almost never holds in this scenario and is cumbersome to test for.

MANCHESTER 1824

The University of Manchester

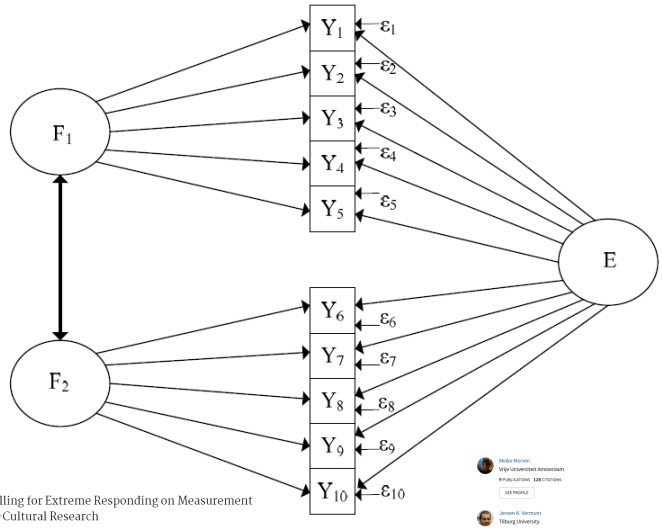
ICLS INTERNATIONAL CENTRE FOR LIFECOURSE STUDIES

Thank you and any questions?
ありがとうございました
質問は?

tarani.chandola@manchester.ac.uk

Collaborators: Benjamin Williams, Noriko Cable

Latent variable model for detecting measurement of response style



The Impact of Controlling for Extreme Responding on Measurement Equivalence in Cross-Cultural Research

Article in Methodology European Journal of Research Methods for the Behavioral and Social Sciences · January 2012