THE ROLE OF MEASUREMENT INVARIANCE IN MEASURES OF MENTAL HEALTH

Nichola Shackleton
Research fellow at COMPASS,
University of Auckland

COMPASSUoA

THE UNIVERSITY OF AUCKLAND
Te Whare Wānanga o Tāmaki Makaurau
NEW ZEALAND

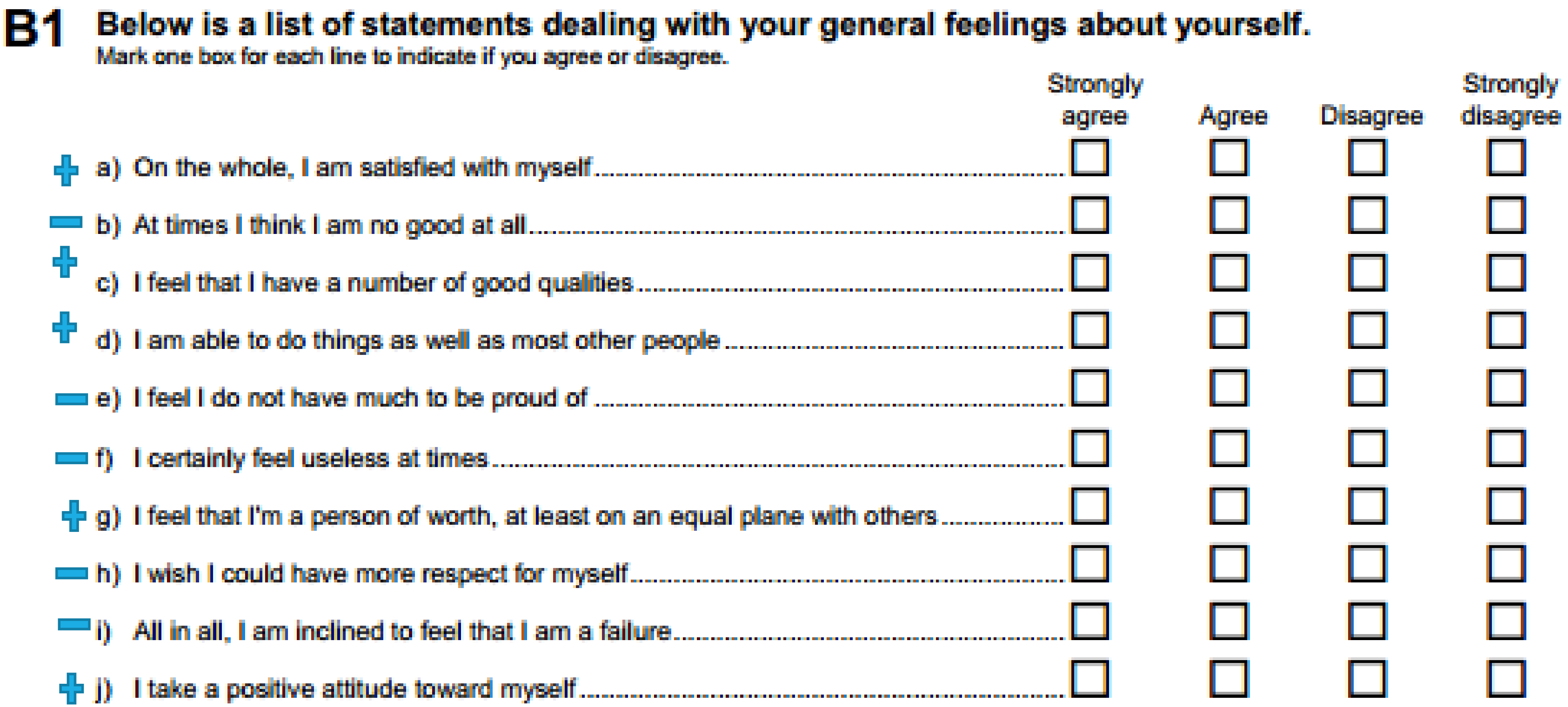
BRITISH ACADEMY
for the humanities and social sciences

BACKGROUND

Research grant looking at inequalities in adolescent substance use and psychosocial health

- Cross country comparisons of inequalities in adolescent substance use
  - Regular smoking, binge drinking, illicit substance use
  - How have these changed over time
- Cross country comparisons of inequalities in adolescent psychosocial health
  - Rosenberg self esteem, Shortened CES-D
  - How have these changed over time

ROSENBERG SELF ESTEEM SCALE

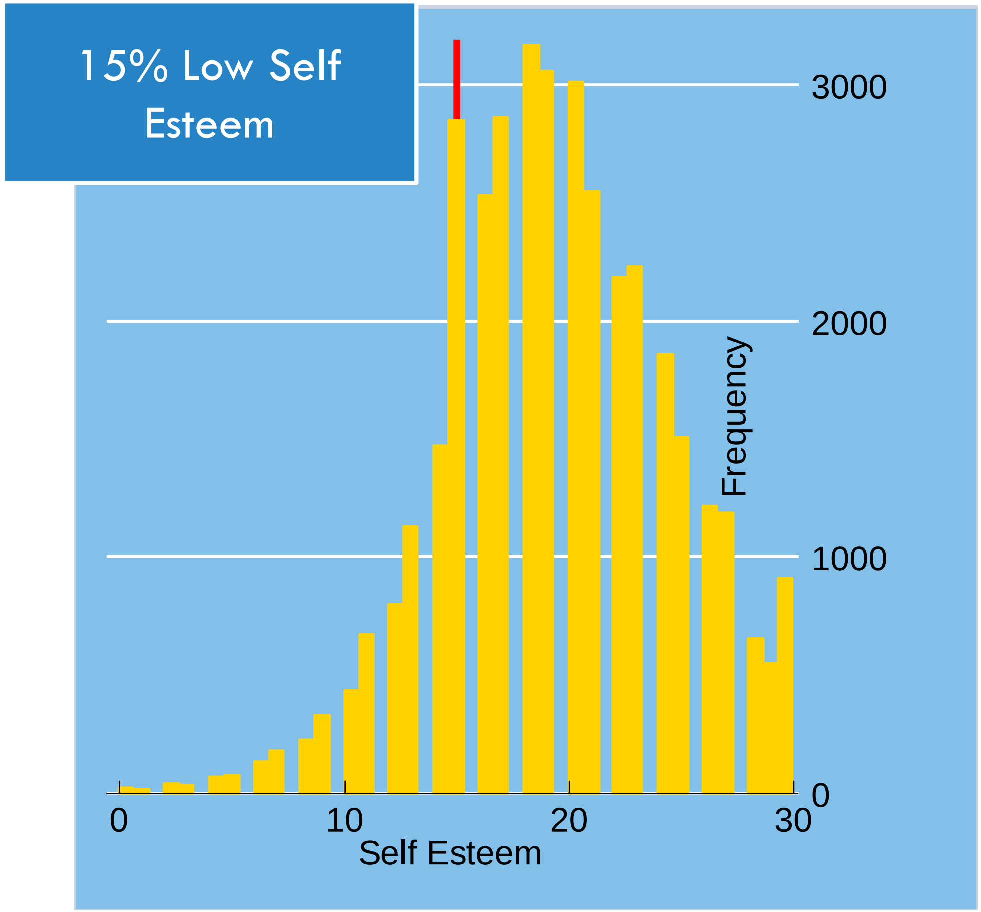


B1 Below is a list of statements dealing with your general feelings about yourself.
Mark one box for each line to indicate if you agree or disagree.

		Strongly agree	Agree	Disagree	Strongly disagree
+	a) On the whole, I am satisfied with myself	□	□	□	□
−	b) At times I think I am no good at all	□	□	□	□
+	c) I feel that I have a number of good qualities	□	□	□	□
+	d) I am able to do things as well as most other people	□	□	□	□
−	e) I feel I do not have much to be proud of	□	□	□	□
−	f) I certainly feel useless at times	□	□	□	□
+	g) I feel that I'm a person of worth, at least on an equal plane with others	□	□	□	□
−	h) I wish I could have more respect for myself	□	□	□	□
−	i) All in all, I am inclined to feel that I am a failure	□	□	□	□
+	j) I take a positive attitude toward myself	□	□	□	□

ROSENBERG SELF ESTEEM SCALE

Country	Mean	SD	N
Slovakia	17.07	3.97	2422
Hungary	17.36	4.62	2762
Faroe Islands	18.01	5.06	543
Isle of Man	18.44	5.44	729
Cyprus	18.53	4.92	6265
Latvia	18.68	4.29	2229
Slovenia	18.81	4.57	3058
Britain	18.87	5.07	2087
Bulgaria	19.05	4.45	2271
Romania	19.31	4.34	2254
Croatia	19.66	4.95	2972
Greece	20.83	5.85	3041
Armenia	21.26	3.97	3928
Iceland	21.31	6.41	3402
Total	19.28	5.07	37963

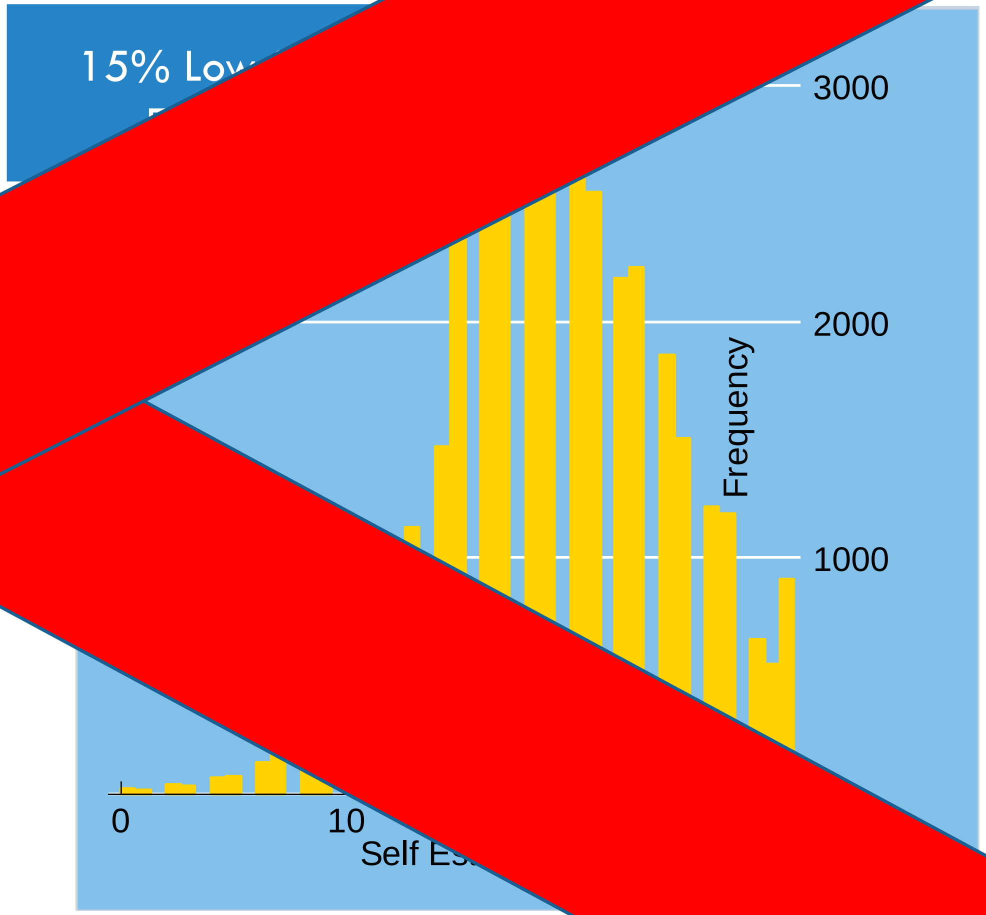

15% Low Self Esteem

ESPAD

The European School Survey Project on Alcohol and Other Drugs

[illegible] SELF ESTEEM SCALE

Country	M[illegible]	[illegible]	N
Slovakia	17.07	[illegible]	[illegible]
Hungary	17.36	[illegible]	[illegible]
Faroe Islands	18.01	5.0[illegible]	[illegible]
Isle of Man	18.44	5.44	[illegible]
Cyprus	18.53	4.92	[illegible]
Latvia	18.68	4.29	[illegible]
Slovenia	18.81	4.57	[illegible]
Britain	18.87	[illegible]	[illegible]
Bulgaria	19.05	[illegible]	[illegible]
Romania	[illegible]	[illegible]	2254
Croati[illegible]	[illegible]	[illegible]	2972
[illegible]	[illegible]	[illegible].85	3041
[illegible]	[illegible]	3.97	3928
[illegible]	[illegible]	6.41	3402
[illegible]	19.28	5.07	37963



ESPAD The European School Su[illegible] and Other Dru[illegible]

MEASUREMENT INVARIANCE

Measurement invariance (or measurement equivalence) is a statistical property of measurement that indicates that the same construct is being measured across some specified groups.

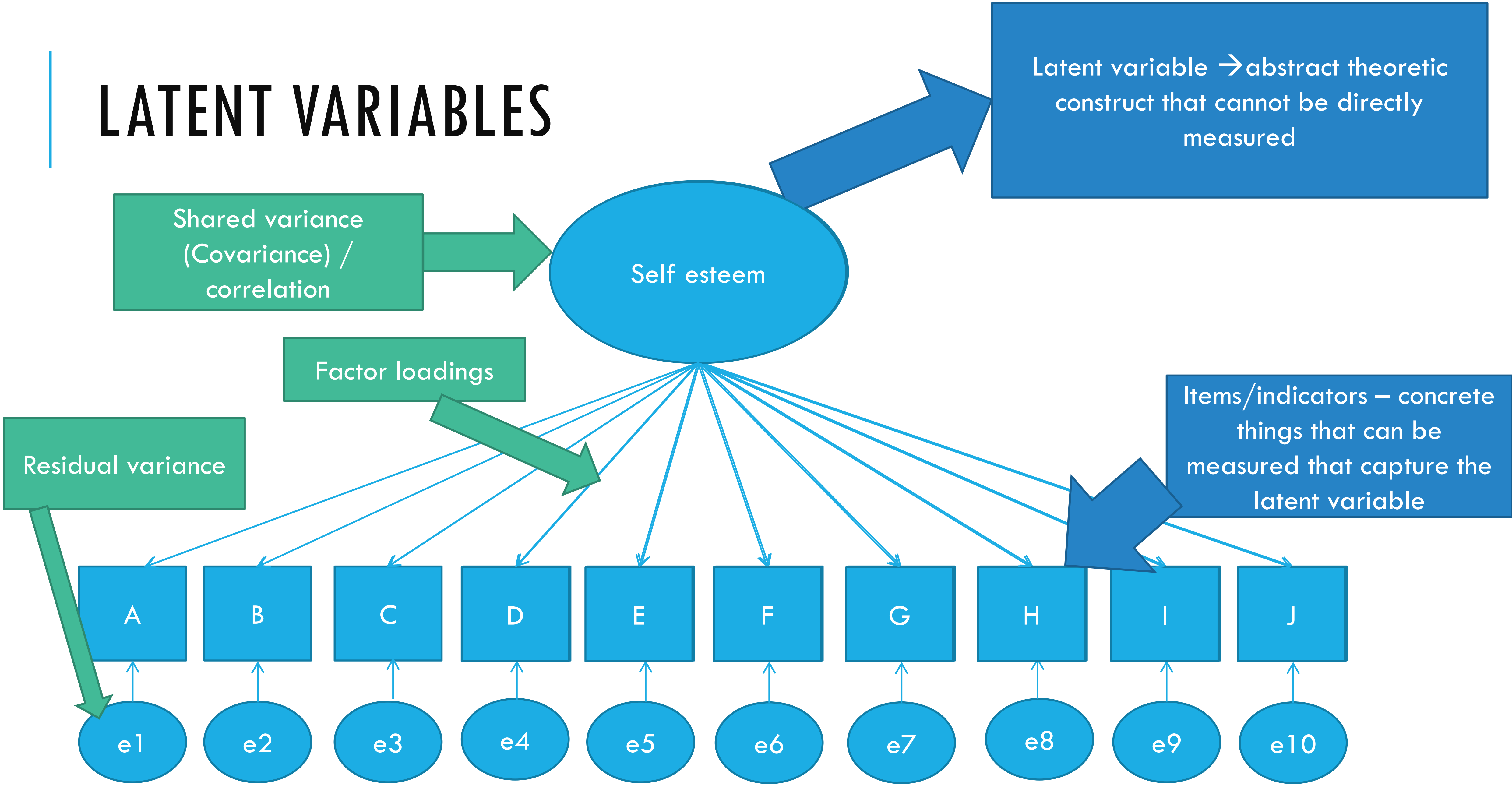
How do we know if the self-esteem scale is measuring the same thing?

Does a mean of 17 in Hungary really reflect a lower average self-esteem than say Armenia (mean=21)?

Does each of the items represent self esteem to the same degree across countries?

Are people with the same level of self esteem in different countries responding to the scale in the same way?

LATENT VARIABLES



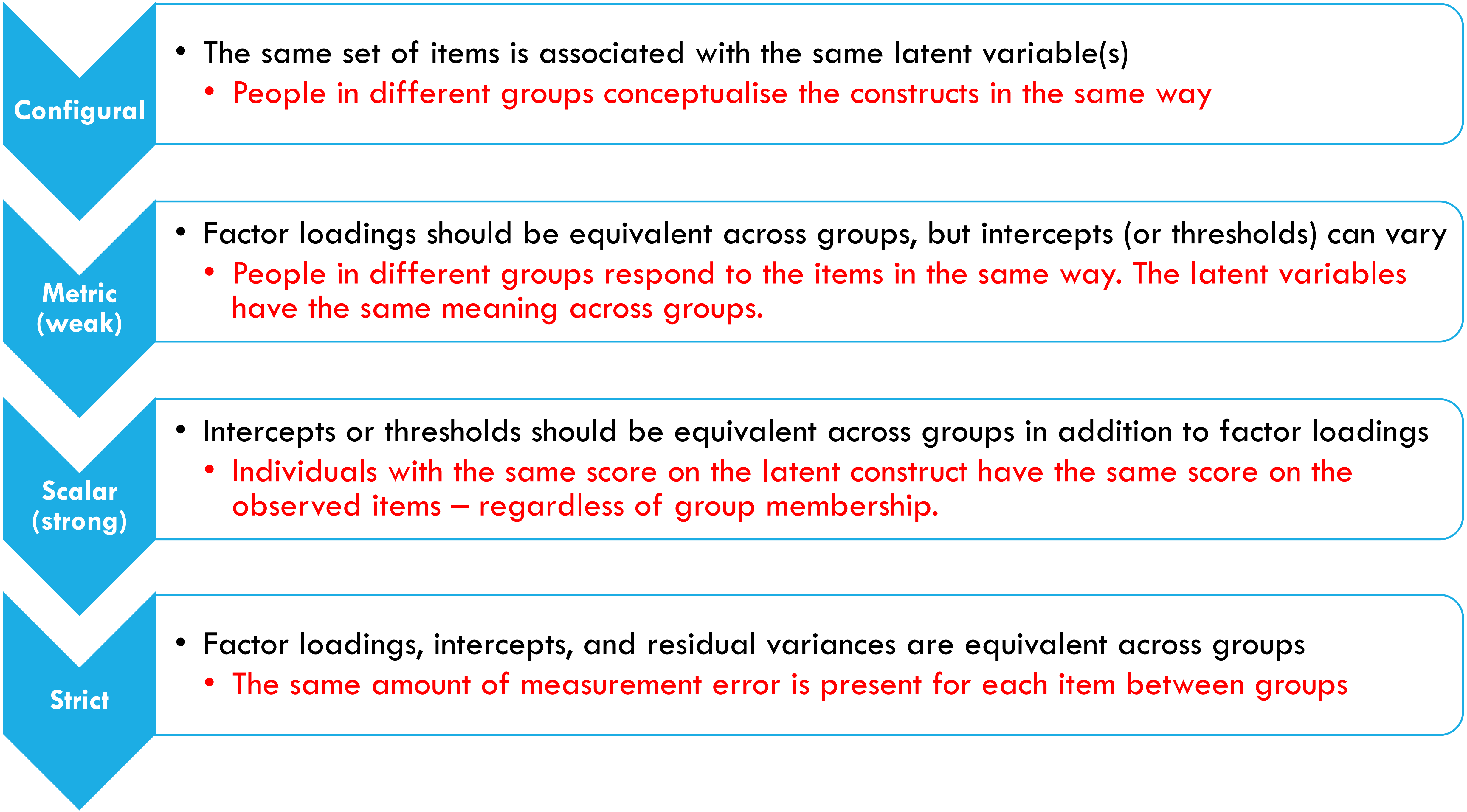
Latent variable →abstract theoretic construct that cannot be directly measured

Shared variance (Covariance) / correlation

Factor loadings

Residual variance

Items/indicators – concrete things that can be measured that capture the latent variable



Configural

- The same set of items is associated with the same latent variable(s)
  - People in different groups conceptualise the constructs in the same way

Metric (weak)

- Factor loadings should be equivalent across groups, but intercepts (or thresholds) can vary
  - People in different groups respond to the items in the same way. The latent variables have the same meaning across groups.

Scalar (strong)

- Intercepts or thresholds should be equivalent across groups in addition to factor loadings
  - Individuals with the same score on the latent construct have the same score on the observed items – regardless of group membership.

Strict

- Factor loadings, intercepts, and residual variances are equivalent across groups
  - The same amount of measurement error is present for each item between groups

Configural

- Test the model fit in each group separately – factorial validity
- Do the groups have the same factor structure?
  - Is the same set of items associated with the same latent variable(s)?
  - Same number of factors, same pattern of loadings?

People in different groups conceptualise the constructs in the same way

CONFIGURAL INVARIANCE → THE SAME SET OF ITEMS IS ASSOCIATED WITH THE SAME LATENT VARIABLE

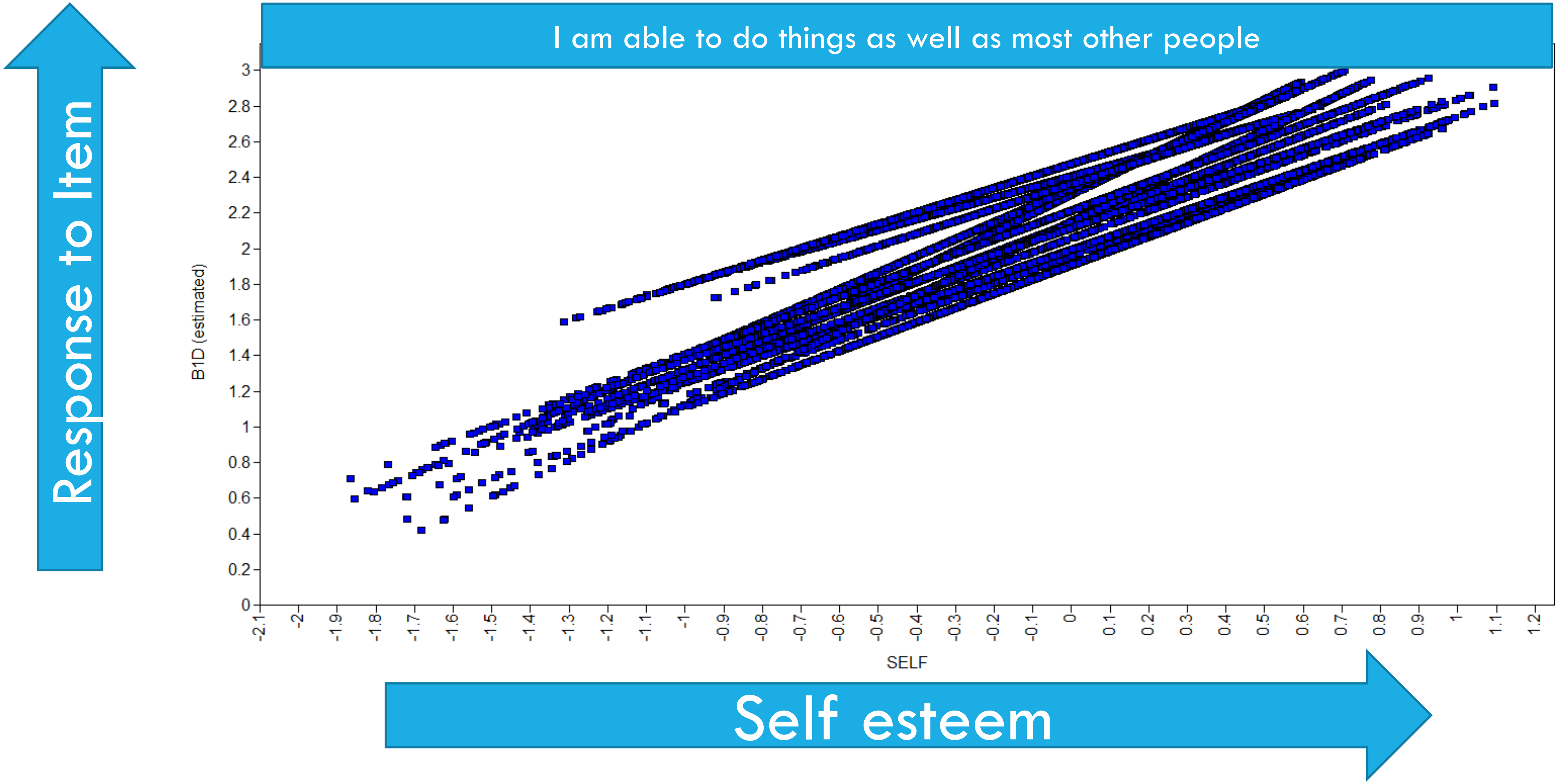
	A	B	C	D	E	F	G	H	I	J
Armenia	0.65	0.02	0.68	0.47	0.05	0.11	0.72	-0.23	0.05	0.70
Bulgaria	0.56	0.64	0.60	0.57	0.55	0.77	0.59	0.36	0.66	0.66
Croatia	0.75	0.75	0.66	0.63	0.64	0.79	0.52	0.47	0.70	0.78
Cyprus	0.68	0.59	0.65	0.60	0.48	0.79	0.62	0.55	0.76	0.78
Faroe Islands	0.81	0.62	0.50	0.66	0.65	0.70	0.69	0.57	0.76	0.87
Greece	0.72	0.80	0.52	0.56	0.45	0.83	0.60	0.50	0.81	0.78
Hungary	0.71	0.49	0.72	0.57	0.35	0.67	0.72	0.33	0.71	0.77
Iceland	0.81	0.80	0.86	0.83	0.61	0.84	0.78	0.69	0.84	0.84
Isle of Man	0.83	0.78	0.81	0.77	0.71	0.79	0.62	0.65	0.76	0.85
Latvia	0.67	0.50	0.75	0.73	0.50	0.48	0.70	0.34	0.58	0.73
Romania	0.53	0.64	0.60	0.54	0.68	0.76	0.52	0.12	0.71	0.50
Slovak Republic	0.64	0.57	0.63	0.55	0.38	0.69	0.66	0.20	0.64	0.70
Slovenia	0.71	0.65	0.71	0.67	0.53	0.74	0.55	0.47	0.71	0.79
United Kingdom	[illegible]	[illegible]	0.77	[illegible]	[illegible]	0.77	[illegible]	[illegible]	[illegible]	[illegible]

Metric

- Constrain factor loadings to be identical across groups
- Does this significantly worsen model fit? Compare to configural model
- reasons for non invariance
  - The meaning differs across groups
  - Some items are more applicable for one group than another
  - Poor translation of scale
  - Groups respond differently to extreme worded items.

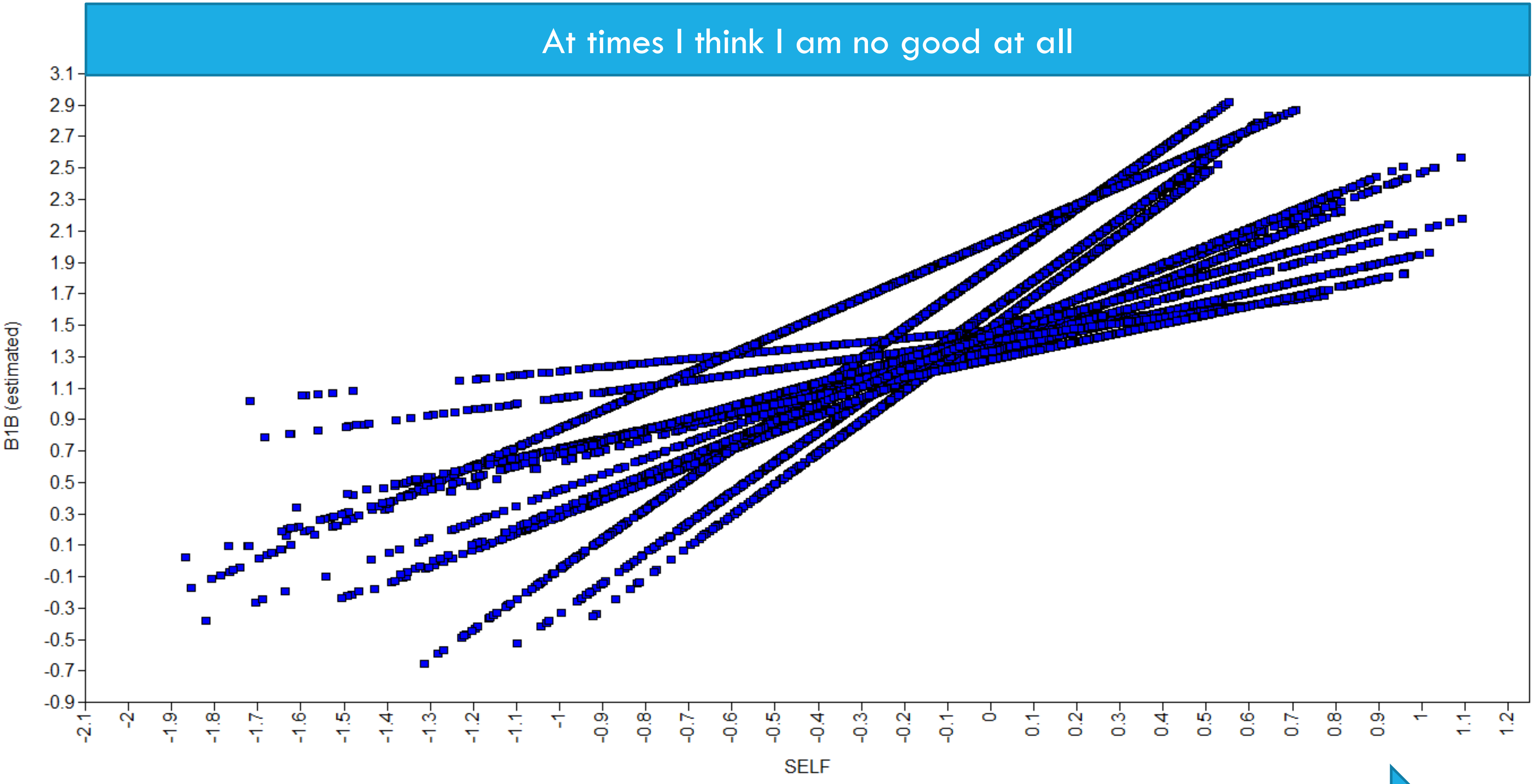
People in different groups respond to the items in the same way. The latent variables have the same meaning across groups.

METRIC INVARIANCE → FACTOR LOADINGS SHOULD BE EQUIVALENT ACROSS GROUPS



METRIC INVARIANCE → FACTOR LOADINGS SHOULD BE EQUIVALENT ACROSS GROUPS

Response to Item


At times I think I am no good at all

Self esteem

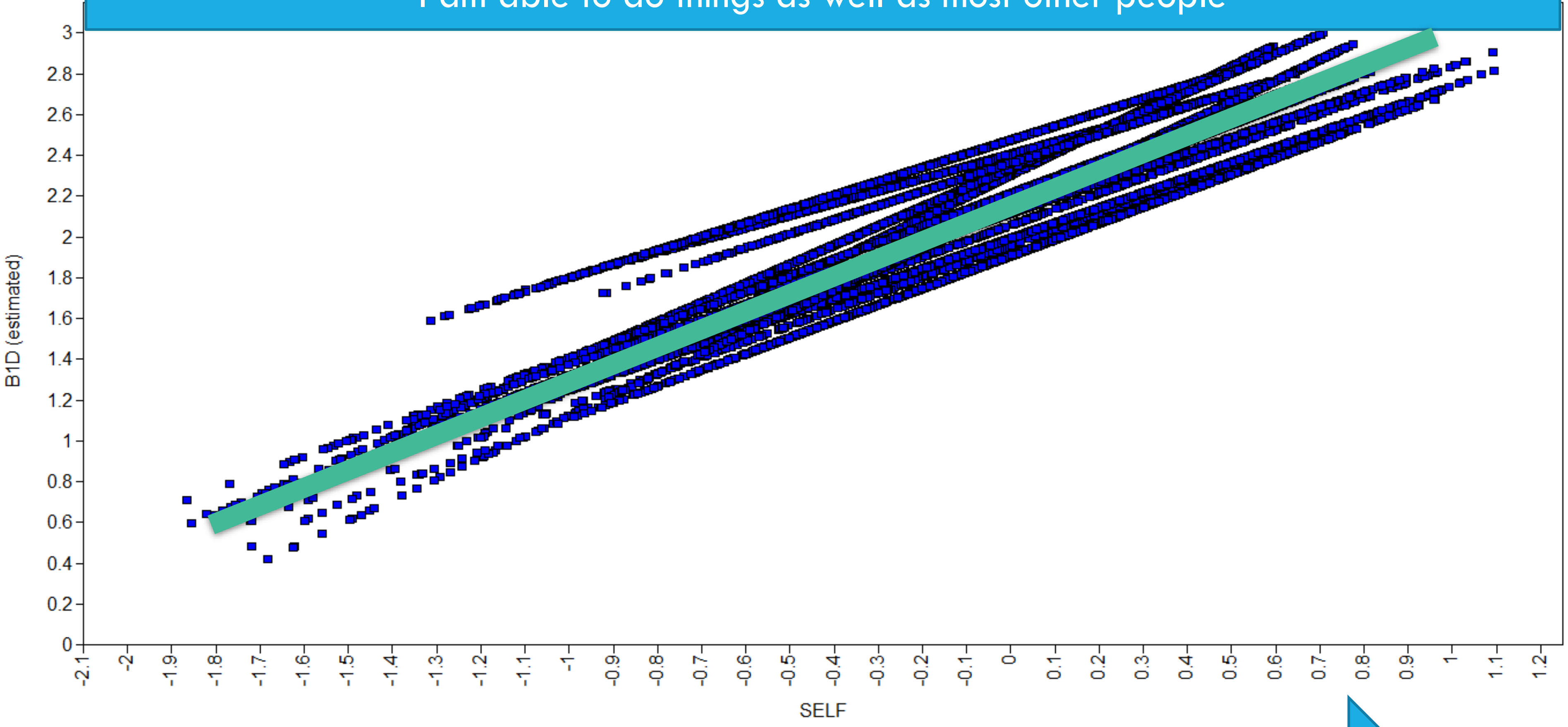
Scalar

- Constrain Intercepts (or thresholds) to be equivalent across groups in addition to factor loadings
- Does this significantly worsen model fit? Compare to metric model
- Reasons for non-invariance :
  - desirability reasons or social norms
  - particular groups displaying a propensity to respond more strongly to an item despite having the same latent trait or factor mean,
  - certain groups having different reference points when making statements about themselves

Individuals with the same score on the latent construct have the same score on the observed items – regardless of group membership

SCALAR INVARIANCE → FACTOR LOADINGS AND INTERCEPTS SHOULD BE EQUIVALENT ACROSS GROUPS

I am able to do things as well as most other people



Response to Item

Self esteem

Strict

- Constrain residual variances to be equal across groups
- Does this significantly worsen model fit? Compare to scalar model
- Do items have the same amount of “not the factor” in them?

ESTABLISHING A CONFIGURAL MODEL

Fit the same model in each group

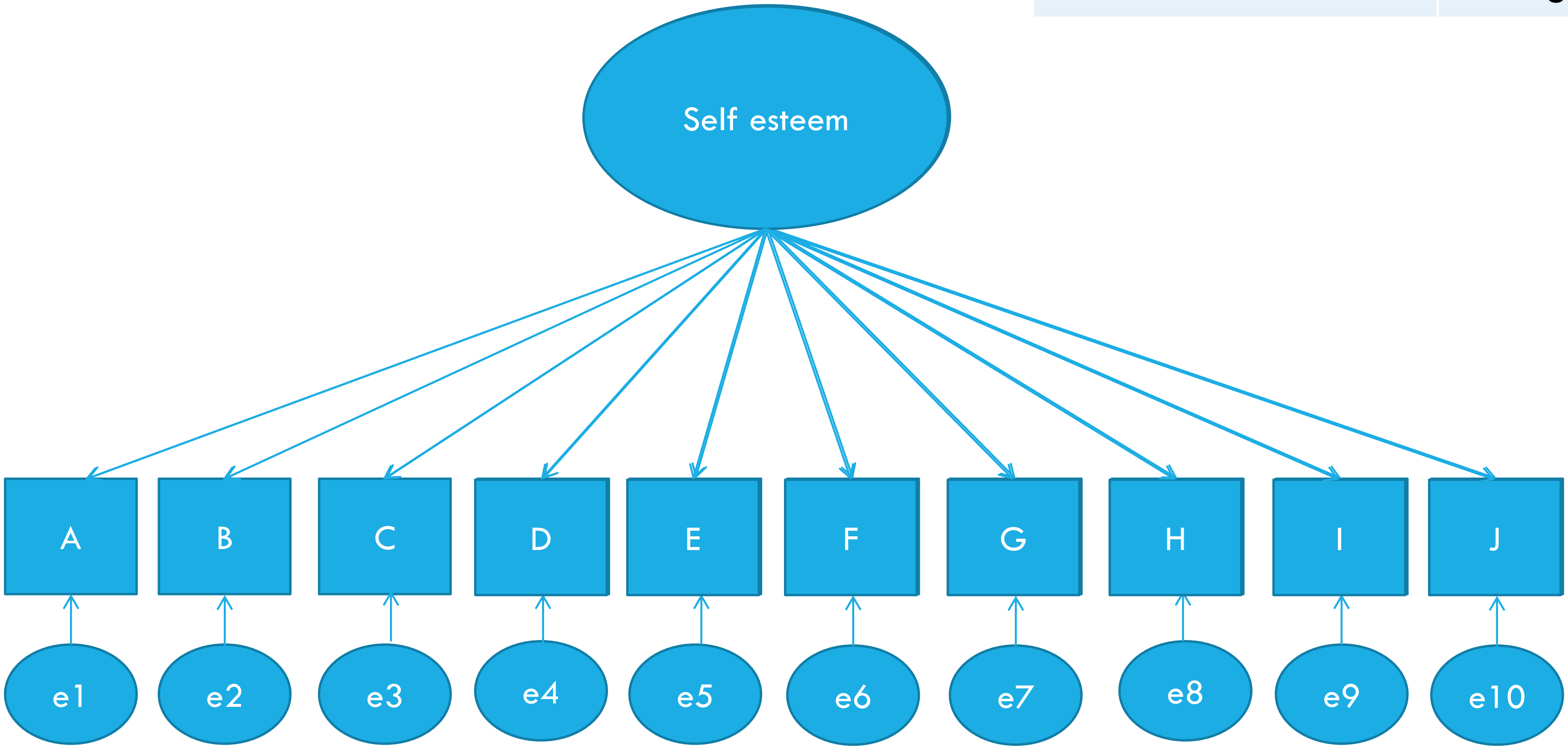
Are the same items correlated with the same factor/factors

Does the correlation structure implied by the data match the correlation structure in the observed data?

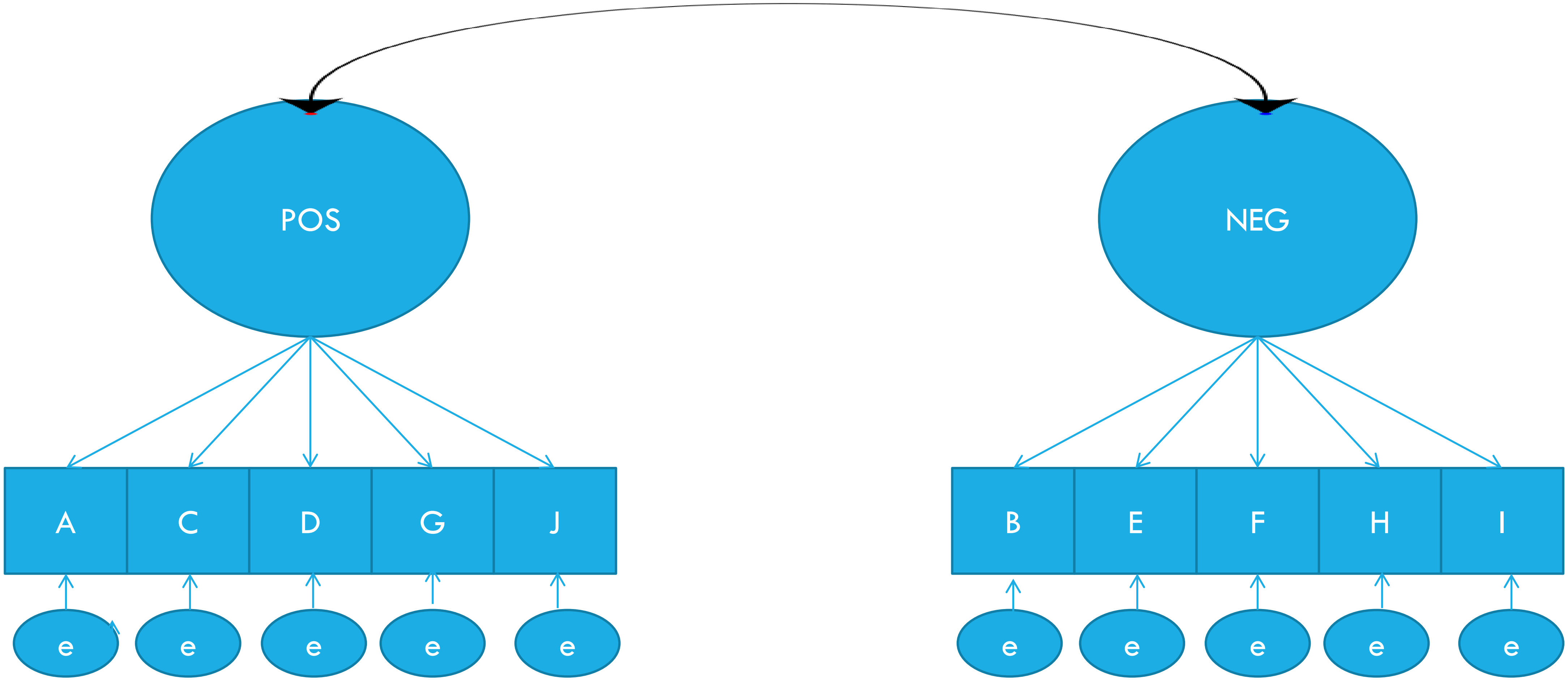
- Model Fit criteria
  - Lower chi squared value
  - CFI > 0.95

THE CONFIGURAL MODEL

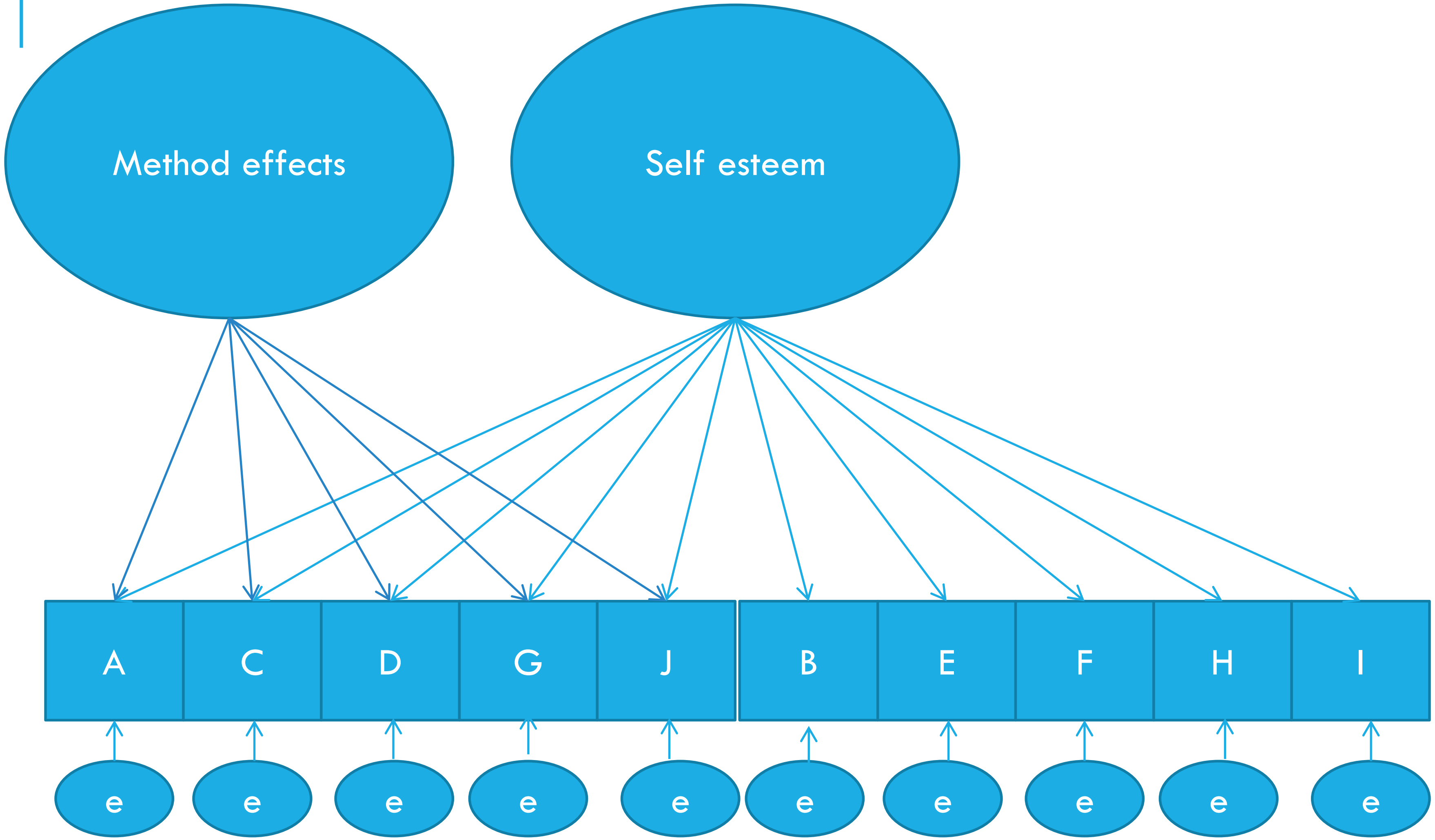
	Configural
χ2(df)	25580.26(351)
CFI	0.664



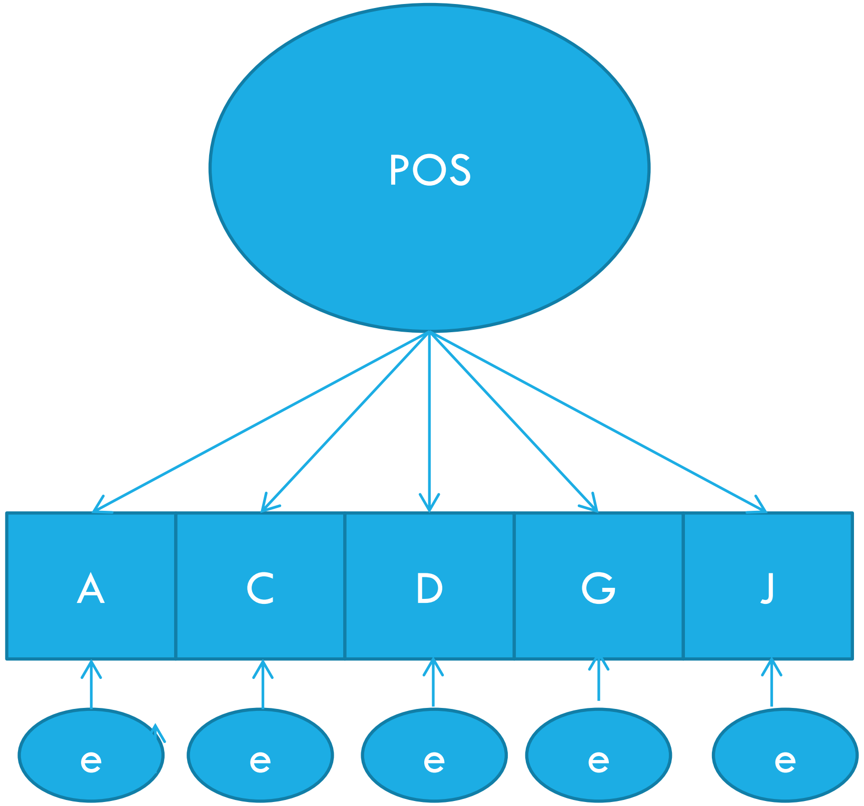
A BETTER CONFIGURAL MODEL?



A BETTER CONFIGURAL MODEL?



A BETTER CONFIGURAL MODEL?



	Configural	Metric	Scalar
χ2(df)	1357.26(765)	1789.95(113)	3622.03(161)
CFI	0.957	0.944	0.885
Metric v configural Δχ2(Δdf)	373.53(48)		
Scalar v metric Δχ2(Δdf)	2018.40(48)		

PARTIAL INVARIANCE

- Scalar invariance is frequently rejected with many groups

Use of modification indices

- Debated criteria
  - At least 2 (total) invariant factor loading/ intercepts/ residual variances

Problematic with many groups

- Many large modification indices – long sequence of modification needed to reach good fit
- Choice of many modification indices can lead to wrong model

ALIGNMENT

```
APPROXIMATE MEASUREMENT INVARIANCE (NONINVARIANCE) FOR GROUPS

 Intercepts/Thresholds
   B1A         2 (3) 4 5 6 (7) (8) 9 10 11 12 (13) 14
   B1C         2 (3) (4) (5) (6) 7 (8) 9 10 11 (12) (13) 14
   B1D         2 (3) 4 (5) (6) 7 (8) 9 (10) (11) (12) (13) 14
   B1G         2 3 (4) 5 (6) (7) (8) 9 10 (11) (12) (13) 14
   B1J         2 3 4 5 (6) 7 8 9 (10) (11) 12 (13) 14

 Loadings for POSSELF
   B1A         2 (3) (4) 5 6 7 8 9 10 (11) 12 13 14
   B1C         2 3 4 (5) 6 7 8 9 10 11 12 13 14
   B1D         2 3 4 5 6 (7) 8 9 10 11 12 13 14
   B1G         2 3 4 5 6 (7) 8 9 10 11 12 13 14
   B1J         (2) 3 4 5 6 7 (8) 9 (10) (11) 12 13 14
```

ALIGNMENT

FACTOR MEAN COMPARISON AT THE 5% SIGNIFICANCE LEVEL IN DESCENDING ORDER

```
 Results for Factor SELF

           Latent     Group       Factor
 Ranking    Class     Value        Mean     Groups With Significantly
                                            Smaller Factor Mean
     1         7         8        0.495     2 3 11 10 13 14 4 9 5 7
                                            12
     2         5         6        0.486     2 3 11 10 13 14 4 9 5 7
                                            12
     3         1         2        0.376     3 11 10 13 14 4 9 5 7 12
     4         2         3        0.214     13 14 4 9 5 7 12
     5        10        11        0.162     13 14 4 9 5 7 12
     6         9        10        0.147     4 9 5 7 12
     7        12        13        0.099     4 5 7 12
     8        13        14        0.081     4 5 7 12
     9         3         4        0.007     7 12
    10         8         9        0.001     7 12
    11         4         5       -0.089     7 12
    12         6         7       -0.292
    13        11        12       -0.301
```

REAL WORLD APPLICATION – INVARIANCE

"Development and community-based validation of eight item banks to assess mental health"

- Item banks to assess mental health issues
- Initial large pool of items tested for local dependence and invariance
- Invariance across age, gender, ethnicity
- IRT analysis identifies which items works best across the continuum of MH

Philip J Batterham and colleagues ANU

CONCLUSIONS

- Invariance is assumed in all analyses but can be explicitly tested with latent variables
- Can’t make straight comparisons across groups (or time) without testing invariance
- Different levels of invariance allow for different types of comparisons
- Can’t assume that a well used measure like the RSES will show good model fit
- In my data set the RSES shows poor fit
  - Known problems with negatively worded items
  - Better fit in all countries with two factors (neg/pos) or models accounting for “method effects” – cross loadings

Questions?

ESPAD The European School Survey Project on Alcohol and Other Drugs

ESPAD

- Multiple waves: 1995, 1999, **2003, 2007**, **2011**
- Over 25 European countries in each year
- 15-16 year old European students
- Sample size ≥ 2400 per country

Compulsory questions: Substance use (alcohol, tobacco, illicit substances)

Optional modules: integration (parental reactions to drug use), **psychosocial health**; deviance, cannabis problems

Model fit statistics

How well the hypothesis model describes the data is measured via model fit. The evaluation of model should be based upon the model as a whole (global model fit), as well as the individual parameters (Byrne, 2011) and should be based upon several model fit criteria (Hooper, Coughlan and Mullen, 2008). In contrast to how a null hypothesis is normally conceptualised in social science, within the SEM framework the null hypothesis is that the specified model holds in the population. The primary focus of the estimation process in SEM is to yield parameter estimates that minimise the discrepancy (the residual) between the sample covariance matrix and population covariance matrix implied by the model (Hu and Bentler, 1999; Little, Slegers and Card, 2006). This objective is achieved by minimizing a discrepancy function (Fmin), where the discrepancy between the sample covariance matrix and the population covariance matrix is least.

To reiterate the above algebraically, we take ‘s’ to represent the sample covariance matrix, ‘∑’ to represent the population covariance matrix, and ‘θ’ to represent a vector of the model parameters, so that ‘∑θ’ represents the covariance matrix implied by the model. The null hypothesis is therefore ‘∑=∑θ’. Fmin reflects the point in the estimation where S- ∑θ = minimum. Fmin therefore measures the extent to which ‘s’ differs from ‘∑θ’. This value is used to calculate Chi Square statistic χ^2, one measure of global model fit (Byrne, 2011; Hu and Bentler, 1999).

Chi Square

The Chi Square statistic represents the discrepancy between the sample covariance matrix, s, and the restricted covariance matrix ∑θ. The formula for the Chi square statistic is shown in equation A.3.

Equation A.3. The chi square statistic.

$$\chi^2 = (N-1)*F_{min}$$

Where χ^2 is the chi square statistic, N is the number observations[45] and F_{min} is the discrepancy function (Hu and Bentler, 1999; Kline, 2011).

Lower values of the chi square statistic indicate a smaller amount of discrepancy between the observed and fitted values. One of the most widely noted disadvantages of the chi square statistic is its sensitivity to sample size (Byrne, 2011a; Hooper, Coughlan and Mullen, 2008; Hu and Bentler, 1999; Kline, 2011). Even very small discrepancies between the sample covariance matrix and the restricted covariance matrix can become highly significant with large sample sizes. However large sample sizes are required in the analysis of covariance structures, and because all models are approximations there will always be some discrepancy.