



MAP ACADEMY

Nebraska Academy for
Methodology, Analytics & Psychometrics

Evaluating Measurement Invariance with Cross Cultural Sensitivity

Leslie R. Hawley, Betty-Jean Usher-Tate,
Sara E. Gonzalez, & Natalie Koziol

Why is this topic important?

- Evaluating Measurement Invariance with Cross Cultural Sensitivity
- No man's an island
 - Diversity
 - Globalization
 - Litigation
 - Professional Standards
 - Score Interpretation & Comparisons

Why is this topic important?

\$\$\$\$\$\$

- School Finance
- Business of testing
- Philanthropists
- Accountability Reforms

Legal Issues

- US Constitution (State Control)
- Federal (ESEA, NCLB, RTTT, ESSE)
- Litigation

Decisions Based on Scores

- High-stakes and low-stakes
- Benefits & Consequences
- Unintended Consequences

Concepts: Validity and Validation

The validity question – *Does the test/instrument do what it is designed to do and do so consistently?*

- *Validity* has a long history in psychology & testing
- *Validity* is assessed through validation research
- *Validation* focuses upon the research that substantiates the evidential basis for test uses
- *The validation process* utilizes both empirical evidence and theoretical bases to support

(Geisinger, in press)

Validation is the responsibility of both the test user (*consumer*) and the test publisher (*vendor*)

2014 Standards (APA, AERA & NCME)

- *Standards for Educational and Psychological Testing (Standards)* defines validity as:
 - “a unitary concept”
 - “the degree to which all accumulated evidence supports the intended interpretation of test scores for the proposed use” (AERA, APA, & NCME, 2014, p. 14)
- The *Standards* considers validity:
 - “the most fundamental consideration in developing and evaluating tests” (AERA, APA, & NCME, 2014, p. 11)
- Validity (property of the scores)
 - Interpretation
 - Decisions (high-stake, low-stake, consequences)
 - Fairness
 - Comparability
 - Trust / Confidence

Validity Argument & Evidence

The validity argument is constructed much like arguments in a court case; there are expectations or standards to uphold and there may also be important mitigating circumstances unique to the measure or sample.

- Where one may look for evidence for a validity argument:
 - Development process
 - Blueprint [definition and outline of construct]
 - Define intended uses for scores
 - Test manual
 - Norm sample
 - Content of the instrument
 - Response processes by test takers
 - Consistency [internal structure of the assessment]
 - Fairness [relationship of test scores with other variables]
 - Outcome impact [benefits and consequences]

Validity as a Unitary Concept

- Historically, validity had been conceptualized categorically: content, construct, discriminant, convergent, . . .
- What may seem like different types of validity are now viewed as different sources of evidence related to the overarching unitary concept of validity
 - “all validity is of one kind, namely, construct validity” (Messick, 1998, p.37)
- Accumulating construct evidence is an umbrella approach that subsumes all validation
 - Includes, but is not limited to:
 - Reliability evidence
 - Statistical conclusion validity evidence
 - Content evidence
 - Convergent, discriminant, and factorial evidence
 - Evaluation of group differences

Validity Evidence

Validation requires a clear argument for the proposed interpretations and uses of scores (Kane, 2006)

- Interpretive argument → inferences from the observed data to any claims we hypothesize
 - Outlines reasoning and provides specific claims that need to be evaluated
 - Framework for evaluation
- Validity argument → evaluation of the interpretive argument

“Validity is an inductive summary of both the existing evidence for and the actual as well as potential consequences of score interpretation and use” (Messick, 1989, p.5)

Validity Evidence

Evidence is based on a particular use and interpretation

- Specific to how we define the construct
- Determines how we can interpret scores from our measure
- Validity is a property of the scores and not the instrument

Evidence should be multifaceted

- Variety of sources and methods
- Need to provide “a convincing, comprehensive validity argument” (Sireci, 2009, p.33)

Validity Evidence

“Multiple lines of evidence . . . consonant with the inference, while establishing that alternative inferences are less well supported” (Messick, 1989, p.5)

Multiple sources for accumulating validity evidence

- Considerations for cultural and linguistic differences
- Test platform and issues of access and/or familiarity
- Today’s focus is primarily on **content** and **construct** sources of evidence

Perception, Trust and Confidence

- Face Validity

- Not always seen as legitimate component of the validity argument
- Empirical methods

Content-related Evidence

Potential questions of interest:

- How well does the measure reflect the intended construct, knowledge, skills?
 - Relevance
 - Representativeness
- How were items developed?
- Were items evaluated prior to administration?
- Were multiple groups (e.g., women, minorities) represented in the development process?

Example of Content-Related Evidence



Pruebas Publicadas en Español

An Index of Spanish Tests in Print

BUROS
CENTER FOR TESTING

Know Yourself

- Cultural Background
- Language
- Language Modality (i.e., Verbal, Nonverbal)
- Education (e.g., level, field)

History: Continuum of Procedures

- Literal translations were [are] standard practice.
 - Forward translation: native speaker of the target language and fluent in the source language.
 - Backward translation: native speaker of the source language and fluent in the target language.
- Societal shifts (i.e., globalization) led to increasing awareness of problems with translations alone.
- Need for adaptations and standardization of procedures arose!

What are the options?

- Literal translation
 - Pro: maintains metric equivalence
 - Con: does not take into account cultural differences; may not be adequate
- Adaptation
 - Pro: Adaptable to specific culture/group
 - Con: Increased difficulty to compare cross-culturally
- New test
 - Pro: Flexible; specific to culture/group
 - Con: Nearly no equivalence maintained

Reasons for Test Adaptation

- Knowledge and skills of interest are often the same across language groups
 - Test adaptation ensures consistency of content
- More efficient than developing a new test
- Test equivalence and fairness is simpler to establish

Steps for Adapting Measures

1. Checking content and format equivalence
2. Decide on suitability of translation/adaptation or creating of new measure/test
3. Select well-qualified translators
4. Translating and adapting process
5. Reviewing the adapted version
6. Conducting a small tryout of the adapted version
7. Carrying out a more ambitious study (check for validity and equivalence –to be discussed later)
8. Document the process

ITC Guidelines

- Documentation of adaptation should be provided, along with evidence of the equivalence.
- Score differences among samples of populations cannot be taken at face value.
 - Researcher has responsibility to verify with other empirical evidence
- Comparisons can only be made at the level of invariance established for the scale.

ITC Guidelines

- Specific information of ways in which the socio-cultural and ecological contexts potentially affects performance should be provided.
 - Test developers should suggest procedures to account for these effects in the interpretation of results.
- Apply appropriate statistical techniques to:
 - establish equivalence of different versions
 - identify problematic components

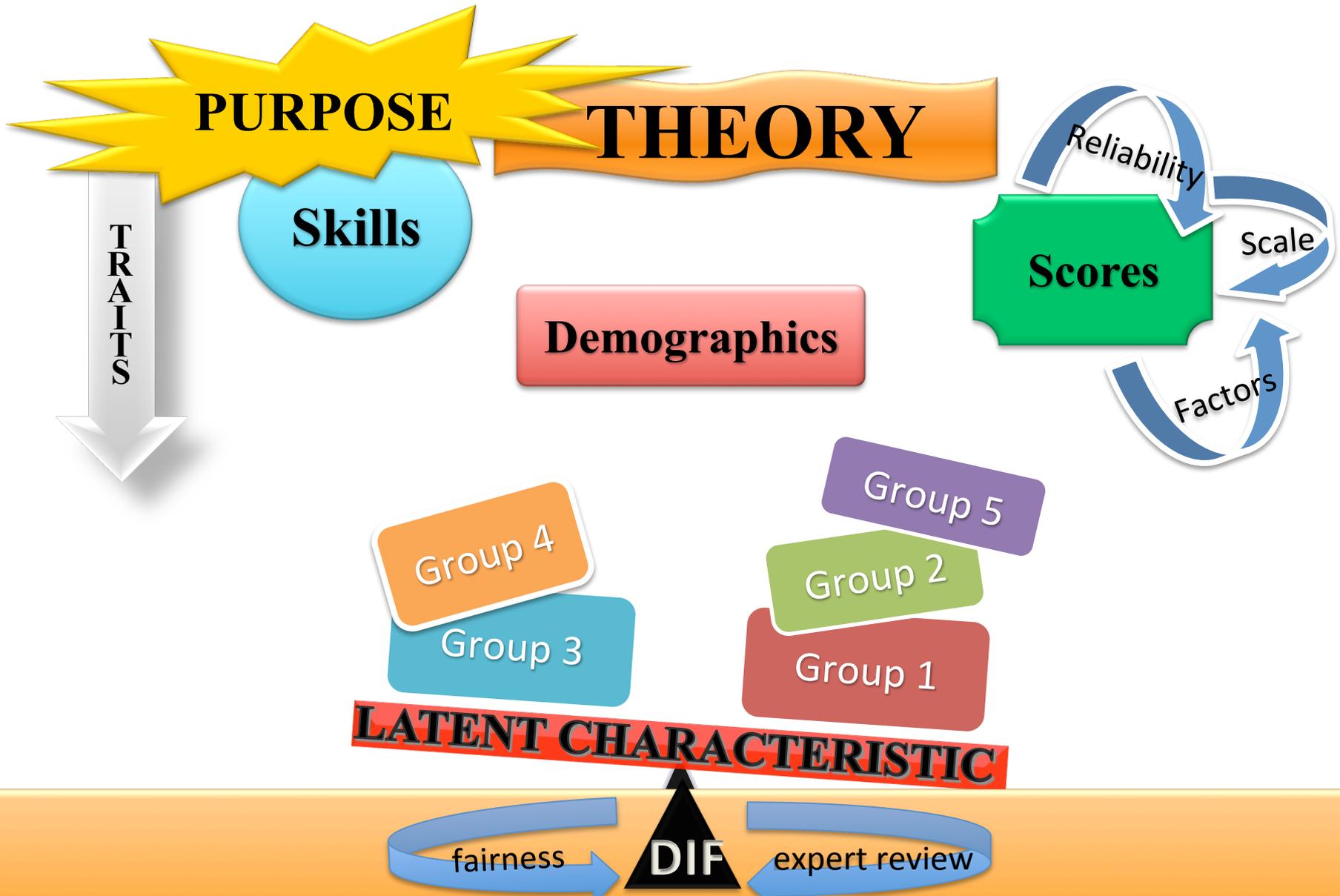
Progress?

- Has progress been made in test adaptation methodology?
- The Buros Center for Testing is changing the way individuals
 - assess their knowledge of testing diverse populations
 - partake in appropriate test selection.

Pruebas Publicadas en Español

- Resource that provides descriptive and analytical information about commercially available tests available in Spanish.
- Material presented in a bilingual manner
- Efforts to point out the need for adaptation
 - availability of norms for Spanish-speaking population
 - country/language the test originated
 - translation or adaptation processes implemented
 - test components
 - original name of the test

CONSTRUCT-RELATED EVIDENCE



Construct-related Evidence

- What seems like different types of validity are different sources of evidence related to the overarching concept of construct validity
 - “all validity is of one kind, namely, construct validity” (Messick, 1998, p. 37)
- Constructs
 - Unobserved, latent characteristics given meaning through the combination of measurable attributes, skills, or traits
 - Ex: Depression, IQ, Conflict, Self-Efficacy, Motivation
 - Operationalization of constructs is guided by theory

Construct-related Evidence

- Construct evidence is based on a particular use and interpretation
 - Specific to how we define our construct
 - Determines how we can interpret scores from an instrument
- For instance, if we want to use a particular instrument to make comparisons between two groups we need to provide evidence of invariance
 - Is my construct measured the same way across groups?

Invariance

- In cross-cultural research we assume that both the instrument and the construct being measured are working the same way across different groups
- We assume the following are equal between groups:
 - Number of factors
 - Pattern of loadings on factors
 - Perception of item content
 - Loading size
 - Item means
 - Construct Dimensionality
 - Relationships between construct dimensions

Invariance

- If our assumptions between groups do not hold then our instrument may not represent the construct equally well across groups and we may not be able to interpret scores from the instrument across groups
- Subsequently, it is important to test the validity of these assumptions

Invariance – Data Example

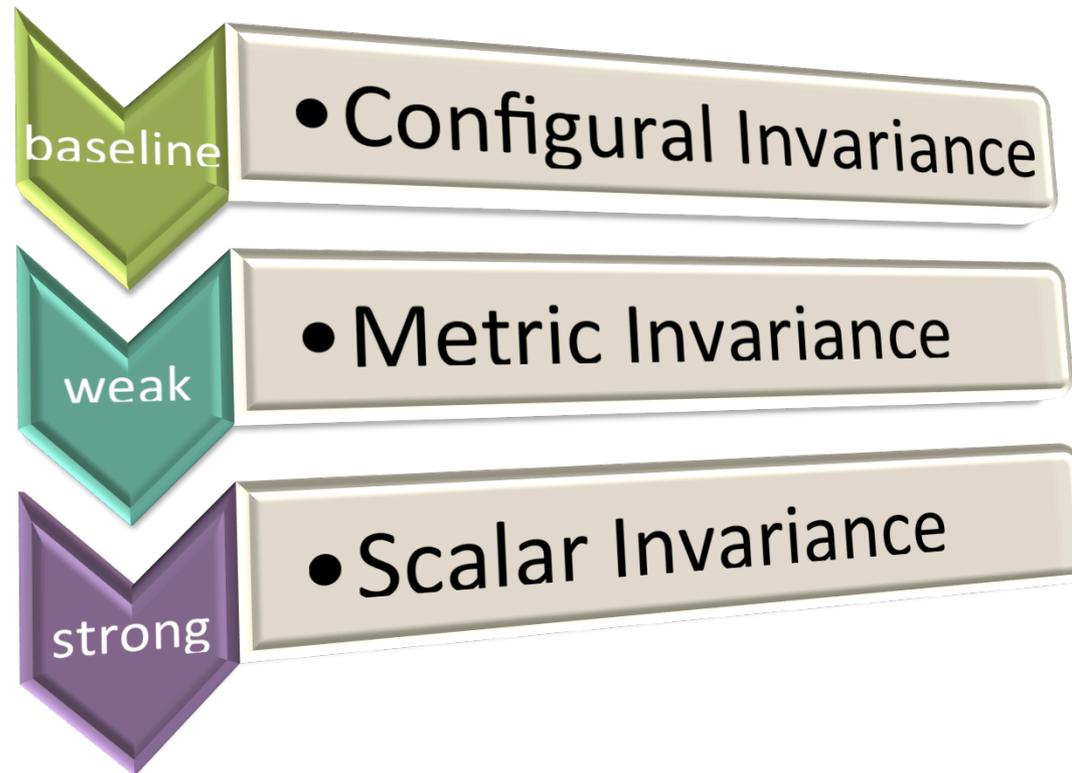
- 2012 Programme for International Student Assessment (PISA)
- 5 item scale: *Teacher Support in Mathematics Classes*
 - “The teacher shows an interest in every student’s learning”
 - “The teacher gives extra help with students need it”
 - “The teacher helps students with their learning”
- Data were collected using complex sampling techniques (students nested within schools)
- Two Countries: USA & Finland

Invariance – Data Example

- Initial analyses attempted to incorporate multilevel structure into invariance testing but the ICCs of the variables were close to 0 (e.g., .05) and models would failed to converge
 - PISA sampling strategy
- Due to multilevel non-convergence, a single level approach was used in the subsequent examples
 - Multiple-Group Confirmatory Factor Analysis (MGCFA)
 - In instances of low ICCs, conventional MGCFA approaches will often provide unbiased estimates (Julian, 2001)

Invariance

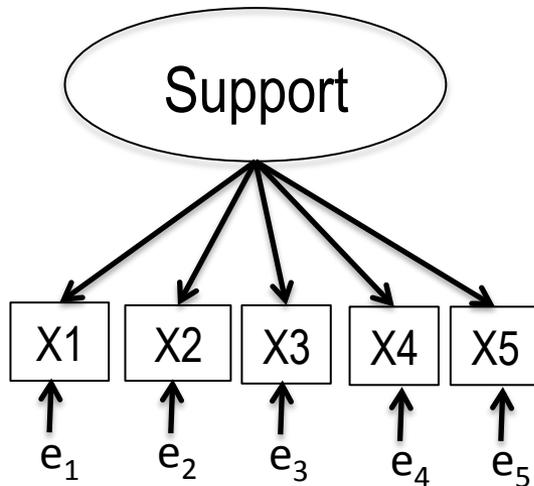
The following steps were conducted to evaluate measurement invariance:



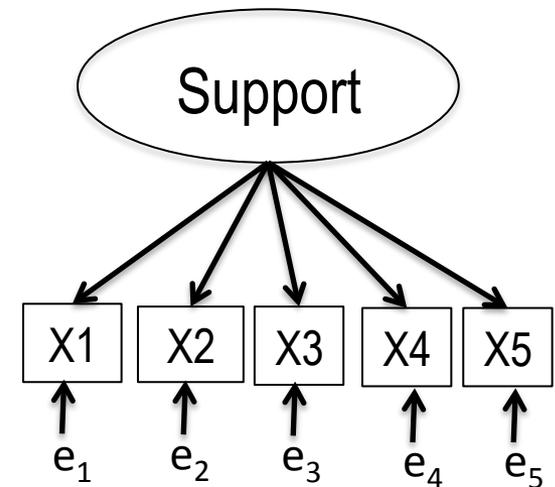
Invariance

- **Configural Invariance (Baseline Model)**
 - Does the same general factor structure (configuration) hold across countries?

United States



Finland



Configural Syntax (Mplus)

- Baseline model
- Everything is separate across groups

Finland (Reference)

```
Model:
!Factor Loadings
Support by
  ST77Q01@1 (L1)
  ST77Q02* (L2)|
  ST77Q04* (L3)
  ST77Q05* (L4)
  ST77Q06* (L5);
ST77Q04 WITH ST77Q02* ;

!Item Intercepts
[ST77Q01*] (I1);
[ST77Q02*] (I2);
[ST77Q04*] (I3);
[ST77Q05*] (I4);
[ST77Q06*] (I5);

!Residual Variances
ST77Q01* (E1);
ST77Q02* (E2);
ST77Q04* (E3);
ST77Q05* (E4);
ST77Q06* (E5);

!Factor Variance;
Support*;

!Factor Mean;
[Support@0];
```

USA

```
Model USA:
!Factor Loadings
Support by
  ST77Q01@1
  ST77Q02*
  ST77Q04*
  ST77Q05*
  ST77Q06* ;

ST77Q04 WITH ST77Q02* ;

!Item Intercepts
[ST77Q01*] ;
[ST77Q02*] ;
[ST77Q04*] ;
[ST77Q05*] ;
[ST77Q06*] ;

!Residual Variances
ST77Q01* ;
ST77Q02* ;
ST77Q04* ;
ST77Q05* ;
ST77Q06* ;

Support*;
[Support@0];
```

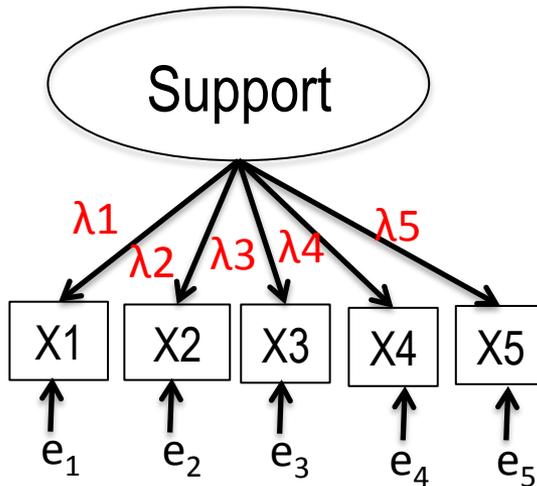
Invariance

- **Metric (weak) Invariance**
 - Do individual items behave similarly across countries?
 - Constraint: Factor loadings (λ) are held equal
 - Partial metric invariance is necessary to make valid inferences in latent factor means (Byrne, Shavelson & Muthén, 1989)

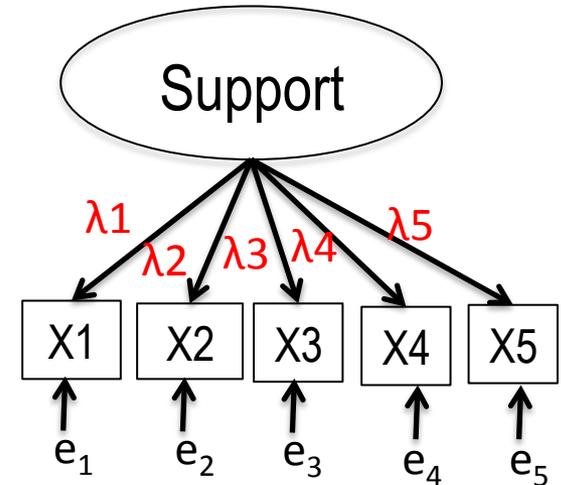
Metric Invariance

- Constraint: Factor loadings (λ) are held equal

United States



Finland



Metric Syntax (Mplus)

- Loadings held equal across groups 
- Factor variance in reference group fixed to 1 

Finland (Reference)

```
Model:
!Factor Loadings
Support by
  ST77Q01@1 (L1)
  ST77Q02* (L2)
  ST77Q04* (L3)
  ST77Q05* (L4)
  ST77Q06* (L5);
ST77Q04 WITH ST77Q02* ;

!Item Intercepts
[ST77Q01*] (I1);
[ST77Q02*] (I2);
[ST77Q04*] (I3);
[ST77Q05*] (I4);
[ST77Q06*] (I5);

!Residual Variances
ST77Q01* (E1);
ST77Q02* (E2);
ST77Q04* (E3);
ST77Q05* (E4);
ST77Q06* (E5);

!Factor Variance;
Support@1;

!Factor Mean;
[Support@0];
```

USA

```
Model USA:
!Loadings held equal to Finland
Support by
  ST77Q01@1 (L1)
  ST77Q02* (L2)
  ST77Q04* (L3)
  ST77Q05* (L4)
  ST77Q06* (L5);
ST77Q04 WITH ST77Q02* ;

!Item Intercepts
[ST77Q01*] ;
[ST77Q02*] ;
[ST77Q04*] ;
[ST77Q05*] ;
[ST77Q06*] ;

!Residual Variances
ST77Q01* ;
ST77Q02* ;
ST77Q04* ;
ST77Q05* ;
ST77Q06* ;

!Factor Variance;
Support* ;

!Factor Mean;
[Support@0];
```

Partial Metric Invariance

- Model fit (i.e., H0 LL; MLR scaling correction) was compared between Configural and Metric
 - Model fit was significantly worse with full metric invariance
 - Modification indices were used to iteratively adjust the model until fit was not significantly worse than the configural model
 - Partial metric invariance was achieved after 2 iterations (only one constraint relaxed at a time)
 - 1 loading was freed
 - 1 residual covariance added for USA only

Invariance

- **Scalar (strong) Invariance**

- Are the meaning of the construct and items equal across countries?

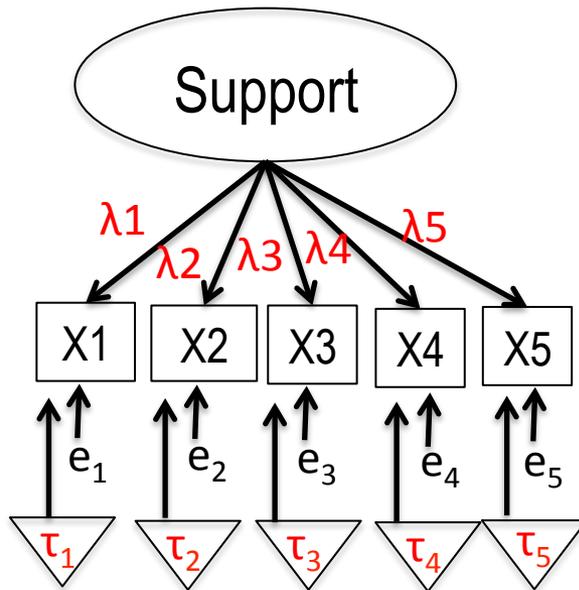
- Constraint: Intercepts (τ) and loadings (λ) held equal

- **Scalar invariance is necessary to compare sum scores or observed means** (van de Schoot, Lugtig & Hox, 2012)

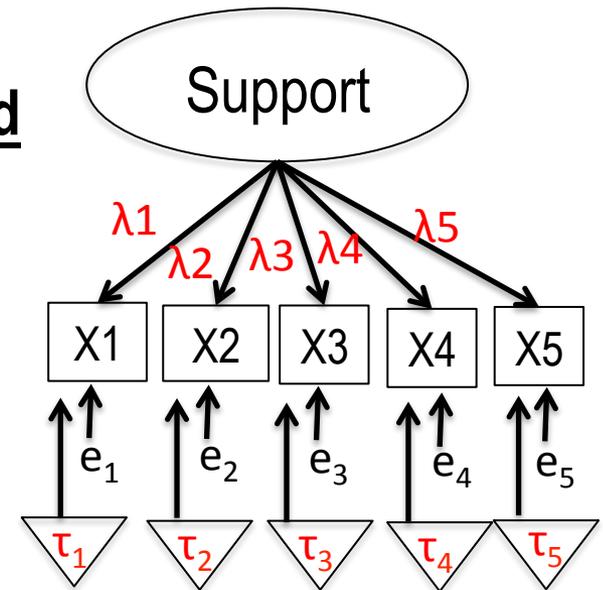
Scalar Invariance

- Constraint: Intercepts (τ) and loadings (λ) held equal

United States



Finland



Partial Scalar Syntax (Mplus)

- Loadings held equal across groups 
- Intercepts held equal across 
- Factor variance in reference group fixed to 1 
- Factor mean of USA now free

Finland (Reference)

```
Model:
!Factor Loadings
Support by
  ST77Q01*
  ST77Q02* (L2)
  ST77Q04* (L3)
  ST77Q05* (L4)
  ST77Q06* (L5);
ST77Q04 WITH ST77Q02* ;

!Item Intercepts
[ST77Q01*] (I1);
[ST77Q02*] (I2);
[ST77Q04*] (I3);
[ST77Q05*] (I4);
[ST77Q06*] (I5);

!Residual Variances
ST77Q01* (E1);
ST77Q02* (E2);
ST77Q04* (E3);
ST77Q05* (E4);
ST77Q06* (E5);

!Factor Variance;
Support@1;

!Factor Mean;
[Support@0];
```

USA

```
Model USA:
!Factor Loadings
Support by
  ST77Q01*
  ST77Q02* (L2)
  ST77Q04* (L3)
  ST77Q05* (L4)
  ST77Q06* (L5);
ST77Q04 WITH ST77Q02* ;
ST77Q01 WITH ST77Q02;

!Item Intercepts
[ST77Q01*] ;
[ST77Q02*] (I2);
[ST77Q04*] (I3);
[ST77Q05*] (I4);
[ST77Q06*] (I5);

!Residual Variances
ST77Q01* ;
ST77Q02* ;
ST77Q04* ;
ST77Q05* ;
ST77Q06* ;

!Factor Variance;
Support*;

!Factor Mean;
[Support*];
```

Partial Metric Invariance

- Model fit (i.e., H0 LL; MLR scaling correction) was compared between conditions
 - Model fit was significantly worse between partial metric and partial scalar conditions
 - Modification indices were used to iteratively adjust the model until fit was not significantly worse than the partial metric model
 - Partial scalar invariance was achieved after 4 iterations
 - 4 intercepts were freed

How did this instrument do?

- Obtained: Partial Scalar invariance
- Minimum Goal: Partial Metric invariance
 - Inferences between latent factor means (Byrne, Shavelson & Muthén, 1989)

How did this instrument do?

- Possible reasons for finding non-invariance
 - Instrument translation
 - Per earlier content discussion
 - Bias (3 types)
 - Construct
 - Differential meanings across groups
 - Method
 - Sample, instrument, administrative
 - Item
 - Content, terminology, unclear wording

What if we have more than 2 groups?

- Limitations of invariance methods with MGCFA and large number of groups
 - Number of groups compared at one time
 - Scalar invariance is rarely achieved with a large number of groups
- Alignment method (Asparouhov & Muthén, 2014)
 - Potential option for multiple groups (up to 100)
 - *Mplus* 7.1
 - Goal is to provide a method for comparing factor means & variances while permitting *approximate* measurement invariance

FINAL THOUGHTS

Reflection



FINAL THOUGHTS

- Validity evidence should be multifaceted
 - Variety of sources and methods
- Evidence is based on a particular use and interpretation
 - Determines how we can interpret scores from our measure
- Cannot ignore cultural components that may influence our constructs
 - Need evidence to demonstrate equality of measurement to interpret scores across groups
- The validation process (accumulation of evidence) is a continual process

FINAL THOUGHTS

- Validity is at the crux for meaningful use of test scores, whether for decisions or comparisons.
- Based on analyses of test reviews published in the Mental Measurement Yearbooks . . . *“favorable evaluations of a test tend to be associated with greater provision of validity evidence.”* (Cizek, Rosenberg, & Koons, 2008)

Questions?
lhawley2@unl.edu