Testing Measurement Invariance Models: Concepts, Applications, and Practical Implementation

2024 National Conference on Educational Measurement and Evaluation
By the Philippine Educational Measurement and Evaluation Association

De La Salle University-Manila || Day 3: August 31, 2024

Johnny T. Amora – *Resource Person*
Director for Institutional Effectiveness and Research
De La Salle-College of Saint Benilde, Manila
johnny.amora@benilde.edu.ph

DE LA SALLE-COLLEGE OF SAINT BENILDE

PEMEA

Agenda

1. **Introduction**
   Physical Measurement Scales vs. Psychological Measurement Scales
   Brief Review of CFA
   Competing CFA Models
   Fitting Data to CFA Models
2. **Measurement Invariance (MI): An Extension of CFA**
   MI Models
   MI Levels
   Practical Guide in Testing MI
   Minimum MI Requirements for One-Group and Multiple-Group Analyses in SEM
3. **Testing MI Using IBM SPSS Amos: A Software Demonstration**
4. **Conclusion**
5. **References**

Measurement Scales

- **Physical Measurement Scales**
  - **Examples:** Tape measure, weighing scale, thermometer, speedometer, etc.
  - **Purpose:** To measure observable physical properties such as length, weight, temperature, or speed.
  - **Characteristics:** Involve a straightforward reading or measurement from a device calibrated to standard units (e.g. kilograms, meters, degrees Celsius).

- **Psychological Measurement Scales**
  - **Examples:** Likert scales, semantic differential scales, IQ Tests, Ability tests, personality inventories.
  - **Purpose:** To measure abstract psychological constructs like attitudes, abilities, traits, emotions, which are not directly observable.
  - **Characteristics:** Composed of multiple items or questions that together quantify a latent variable.

Measurement Model (CFA Model) Example

- For discussion purposes, let's use the **Mental Ability Test** described in Holzinger & Swineford (1939).

Factors	Indicators/Subtests	Description
Spatial	**Visperc**	Measures the ability to perceive spatial relationships.
	Cubes	Assesses spatial reasoning and manipulation of objects.
	Lozenges	Tests the ability to recognize and match shapes.
Verbal	**Paragraph**	Evaluates comprehension of written text.
	Sentence	Measures the ability to understand and construct sentences.
	Word Meaning	Tests vocabulary knowledge and word comprehension.

Additional Variable: Gender

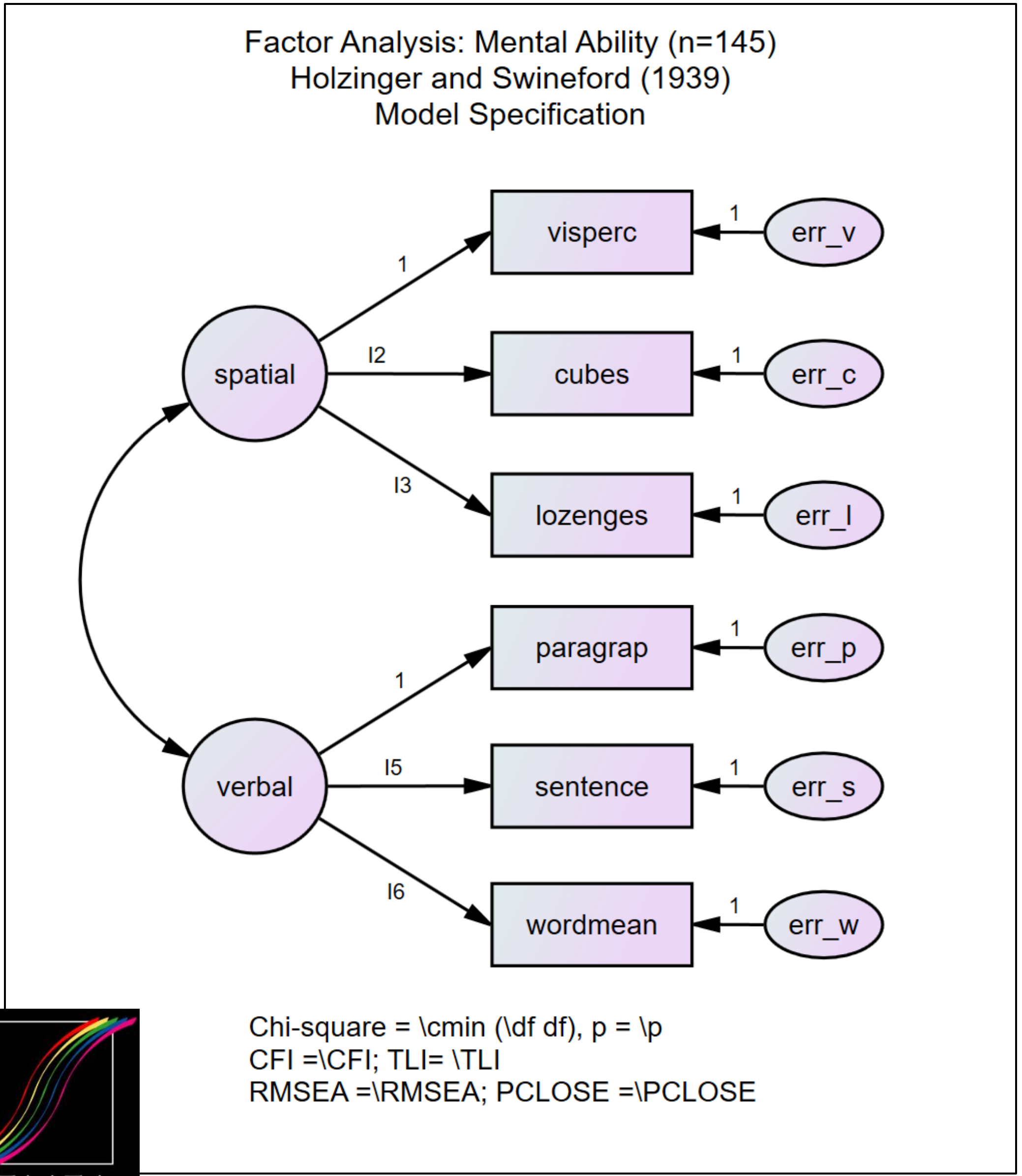
When conducting a Confirmatory Factor Analysis, the data may be fitted to the following competing CFA models:

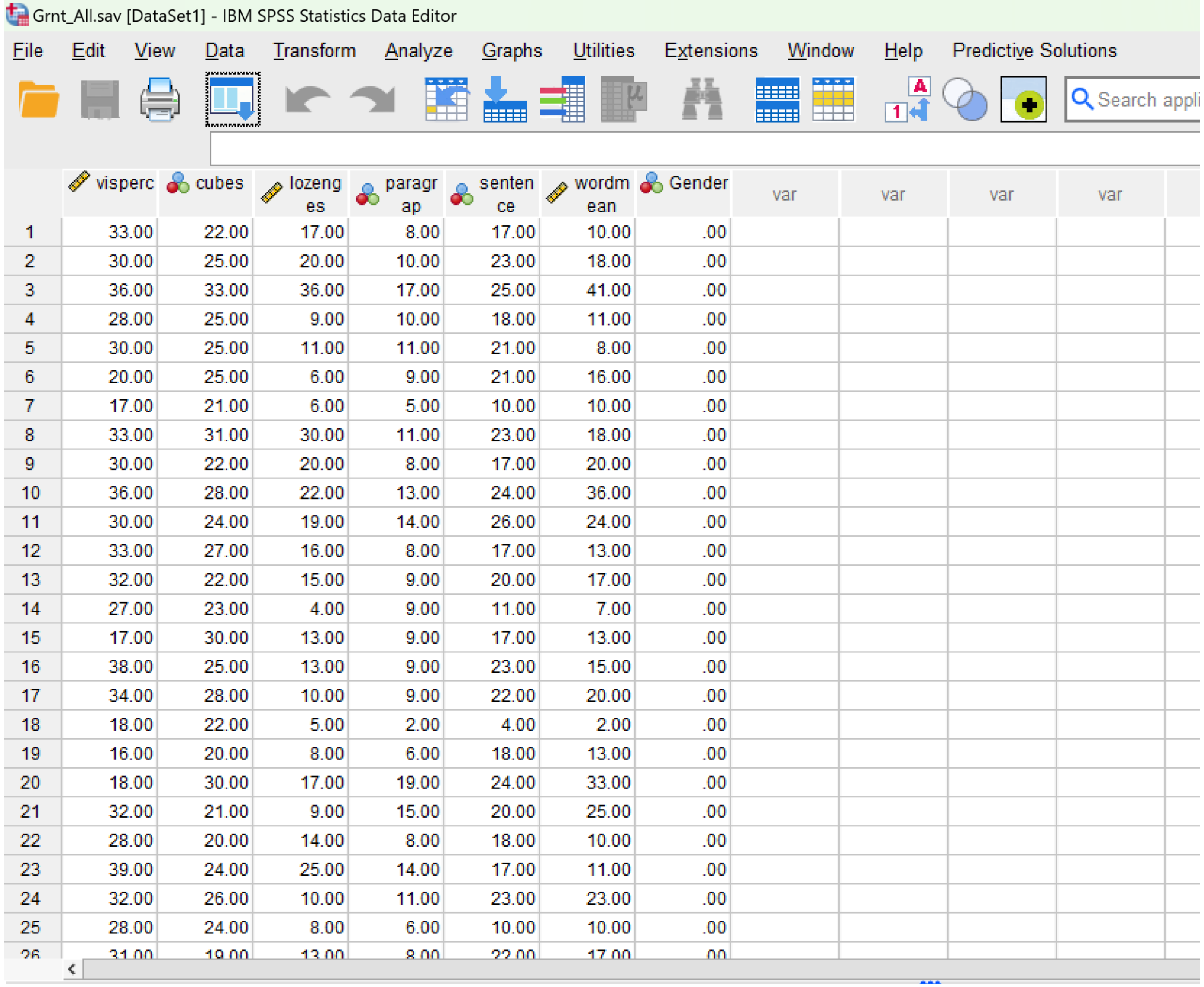
- ✓ **Single Factor Model**
- ✓ **Correlated Factor Model**
- ✓ **Non-Correlated Factor Model**
- ✓ **Second-order Factor Model (Hierarchical Factor Model)**
- ✓ **Bifactor Model**

In this lecture, only the correlated factor model is used as an example.

Correlated Factor Model

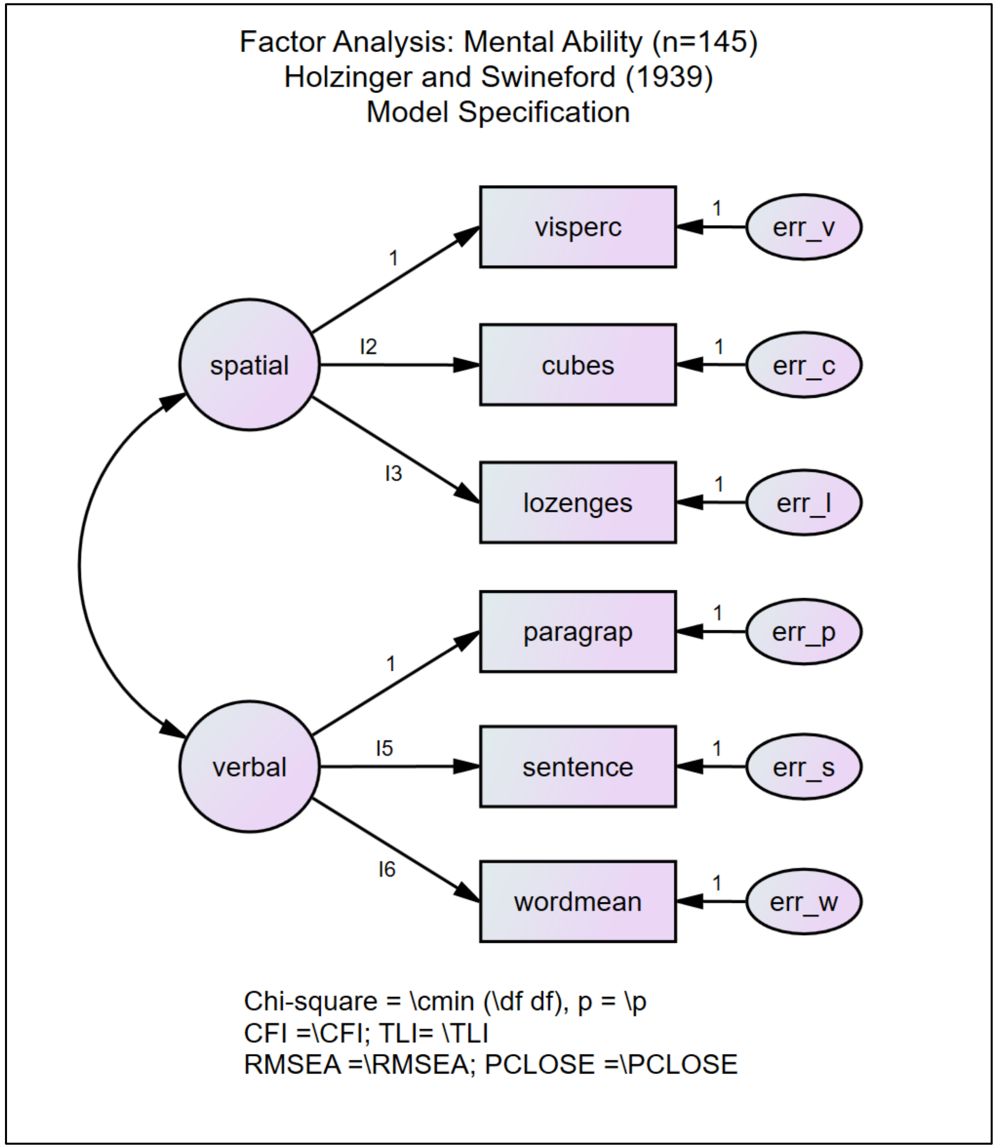
- **Dataset: H**olzinger & Swineford's (1939) dataset available in the PSIMAGO Pro (IBM SPSS Manual).



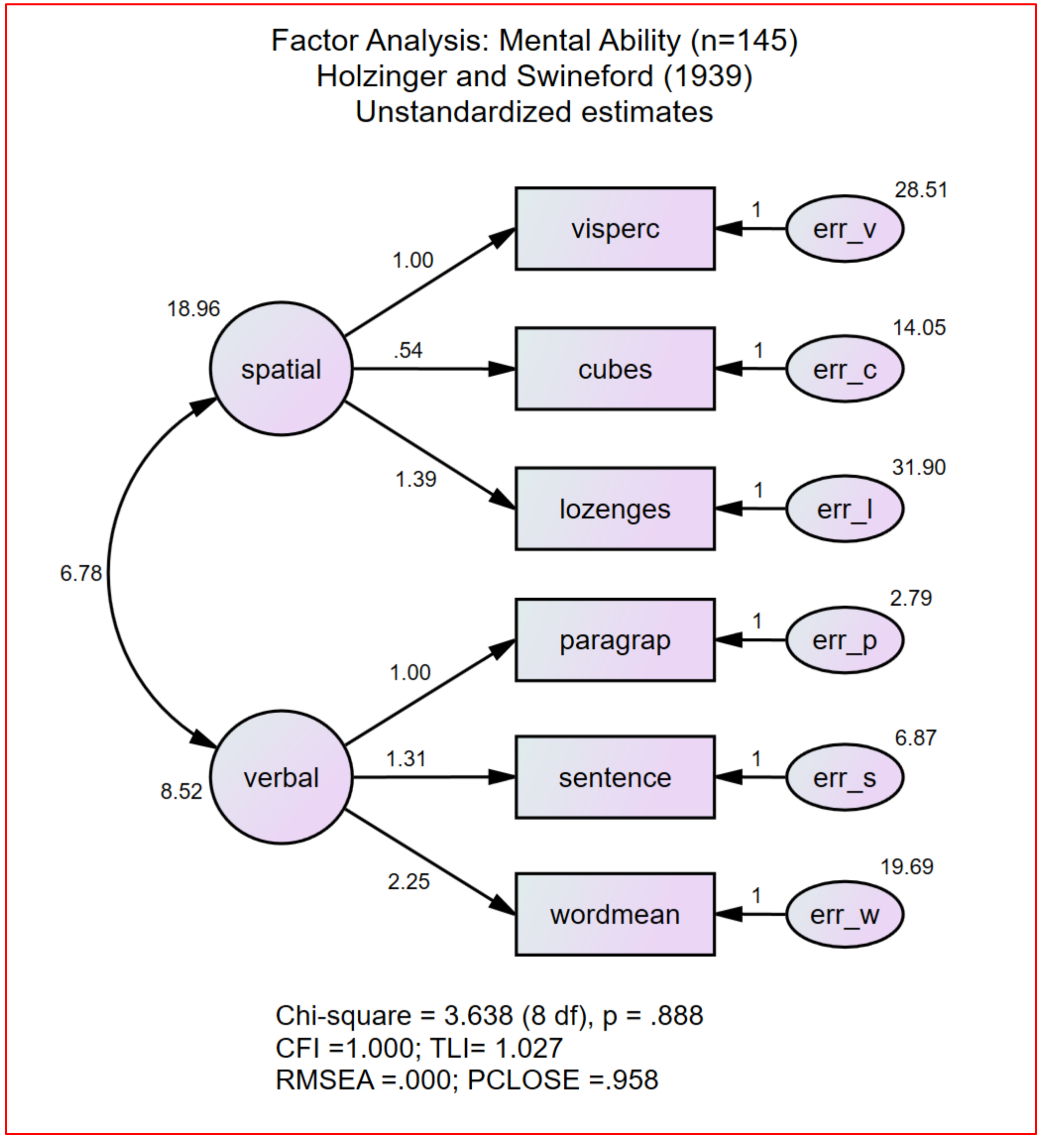


	visperc	cubes	lozenges	paragrap	sentence	wordmean	Gender	var	var	var	var
1	33.00	22.00	17.00	8.00	17.00	10.00	.00				
2	30.00	25.00	20.00	10.00	23.00	18.00	.00				
3	36.00	33.00	36.00	17.00	25.00	41.00	.00				
4	28.00	25.00	9.00	10.00	18.00	11.00	.00				
5	30.00	25.00	11.00	11.00	21.00	8.00	.00				
6	20.00	25.00	6.00	9.00	21.00	16.00	.00				
7	17.00	21.00	6.00	5.00	10.00	10.00	.00				
8	33.00	31.00	30.00	11.00	23.00	18.00	.00				
9	30.00	22.00	20.00	8.00	17.00	20.00	.00				
10	36.00	28.00	22.00	13.00	24.00	36.00	.00				
11	30.00	24.00	19.00	14.00	26.00	24.00	.00				
12	33.00	27.00	16.00	8.00	17.00	13.00	.00				
13	32.00	22.00	15.00	9.00	20.00	17.00	.00				
14	27.00	23.00	4.00	9.00	11.00	7.00	.00				
15	17.00	30.00	13.00	9.00	17.00	13.00	.00				
16	38.00	25.00	13.00	9.00	23.00	15.00	.00				
17	34.00	28.00	10.00	9.00	22.00	20.00	.00				
18	18.00	22.00	5.00	2.00	4.00	2.00	.00				
19	16.00	20.00	8.00	6.00	18.00	13.00	.00				
20	18.00	30.00	17.00	19.00	24.00	33.00	.00				
21	32.00	21.00	9.00	15.00	20.00	25.00	.00				
22	28.00	20.00	14.00	8.00	18.00	10.00	.00				
23	39.00	24.00	25.00	14.00	17.00	11.00	.00				
24	32.00	26.00	10.00	11.00	23.00	23.00	.00				
25	28.00	24.00	8.00	6.00	10.00	10.00	.00				
26	31.00	19.00	13.00	8.00	22.00	17.00	.00				

CFA Model without coefficients



CFA Model with coefficients



Confirmatory Factor Analysis

- **In CFA**, we begin with a theory or CFA model, gather data, and then use that data to test the model.
- **In testing the CFA model**, the factor loadings should be statistically significant ($p<.05$), and the model should demonstrate a good fit with the data (e.g., chi-square $p > .05$). Additionally, the convergent validity, discriminant validity, and reliability should be established.

Measurement Invariance

Measurement Invariance is an extension to CFA, involving the testing of:

- ✓ **Configural Invariance**
- ✓ **Metric Invariance**
- ✓ **Scalar Invariance**
- ✓ **Strict Invariance**

Measurement Invariance

- **Variance:** Indicates difference or variation.
- **Invariance:** indicates consistency or equivalence.
- **Measurement invariance:**
  - **Measurement model** functions consistently across different groups or time points, enabling valid comparisons.
  - Equivalence of factor structure, loadings, intercepts, and residuals across groups or time points.
  - When measurement invariance is established, the researchers can confidently compare scores, ensuring the same construct is measured uniformly across groups.

Pioneers of Measurement Invariance

- **Jöreskog (1971)** was the first to discuss measurement invariance, initially referring to it as the **equivalence of factor structures**.
- **Byrne, Shavelson, and Muthén (1989)** introduced the concept of Measurement Invariance, after which the testing of MI took off.

Jöreskog, K. G. (1971). Simultaneous factor analysis in several populations. **Psychometrika, 36**(4), 409-426. https://doi.org/10.1007/BF02291366
Byrne, B. M., Shavelson, R. J., & Muthén, B. (1989). Testing for the equivalence of factor covariance and mean structures: The issue of partial measurement invariance. **Psychological Bulletin, 105**(3), 456-466. https://doi.org/10.1037/0033-2909.105.3.456

The Configural Invariance

Weighing Scale Analogy	Psychological Measurement Scale
Same Setup: The weighing scales used for different groups or at different time points should have the same structure and operate in the same manner (e.g., digital or mechanical). **Same Components:** The scales used for different groups or at different time points should have identical components (e.g., a platform for placing the object, a dial or digital display).	**Ensuring the scale setup is consistent across groups or over time points:** The same construct is measured using identical items and factor structure across groups or over different time points.

- **Measuring Weights over different time points:** Use the same or identical scale to ensure consistent set up and components.

- **Measuring weights of groups:** Ensure the scales used for the groups have identical set up and components.

Steps in Testing Configural Invariance:

1. **Fit the CFA Model to the Overall Data:**
   - Ensure that the CFA model has an acceptable fit when applied to the overall dataset. This serves as a prerequisite for further invariance testing.
2. **Fit the CFA Model Separately for Each Subgroup or Time Point:**
   - Perform a separate CFA for each subgroup or time point. Although the CFA is conducted separately for each group, there should be only one set of model fit statistics that is acceptable across all subgroups, indicating consistent factor structure.

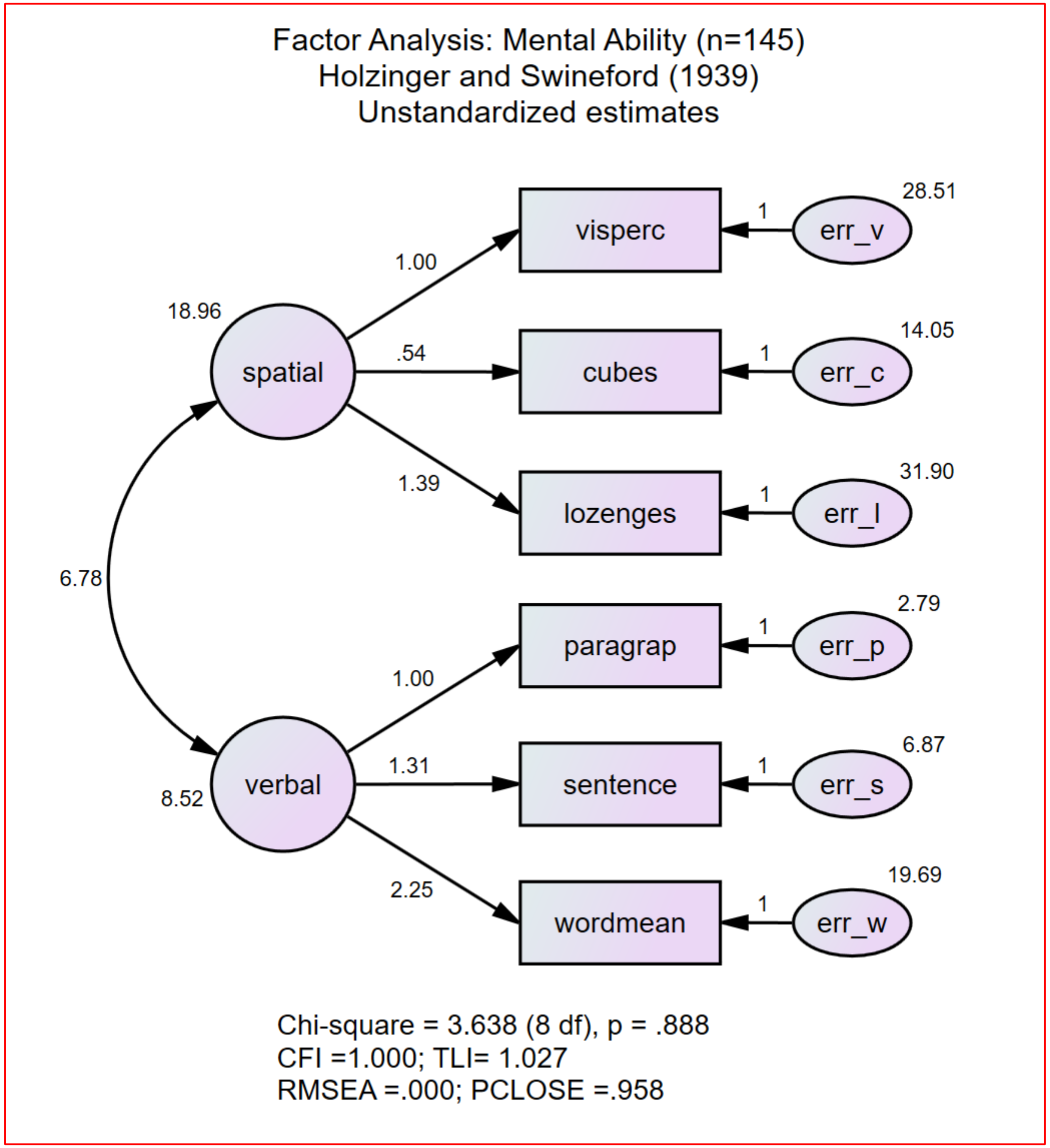
Configural Invariance holds if the following conditions are met:

- The CFA model fits well with the overall data (this is a prerequisite).
- The same CFA model fits well across subgroups or different time points when tested separately.

Example: Testing Configural Invariance

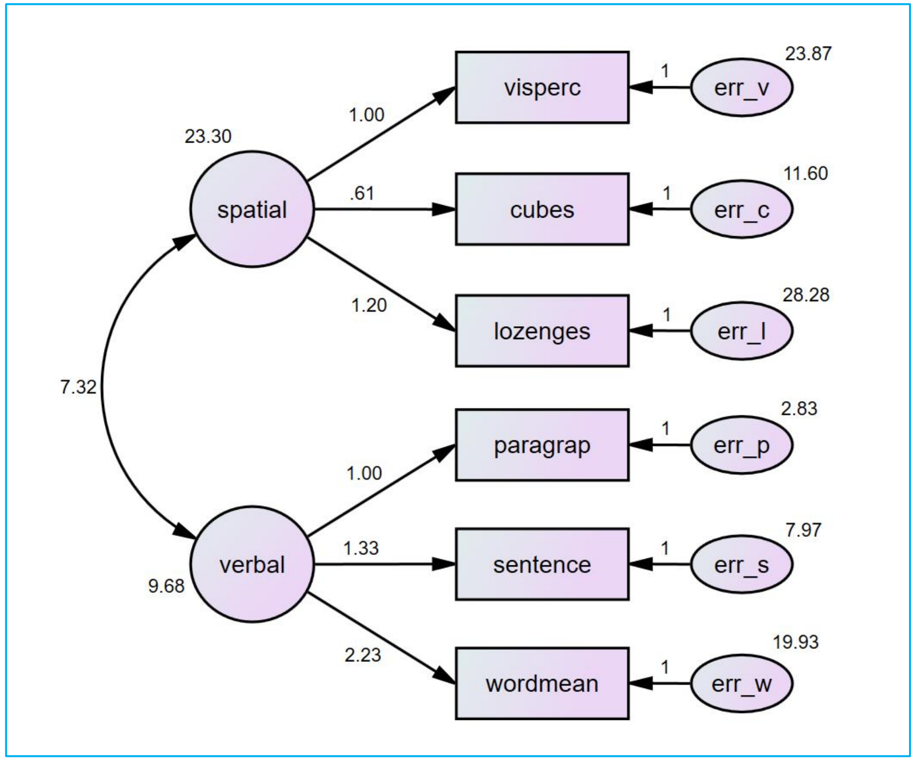
- **Step 1: Fitting the CFA Model to the Overall Data**

Note: Text outputs are not shown here, but the results are satisfactory (e.g., all factor loadings are statistically significant).

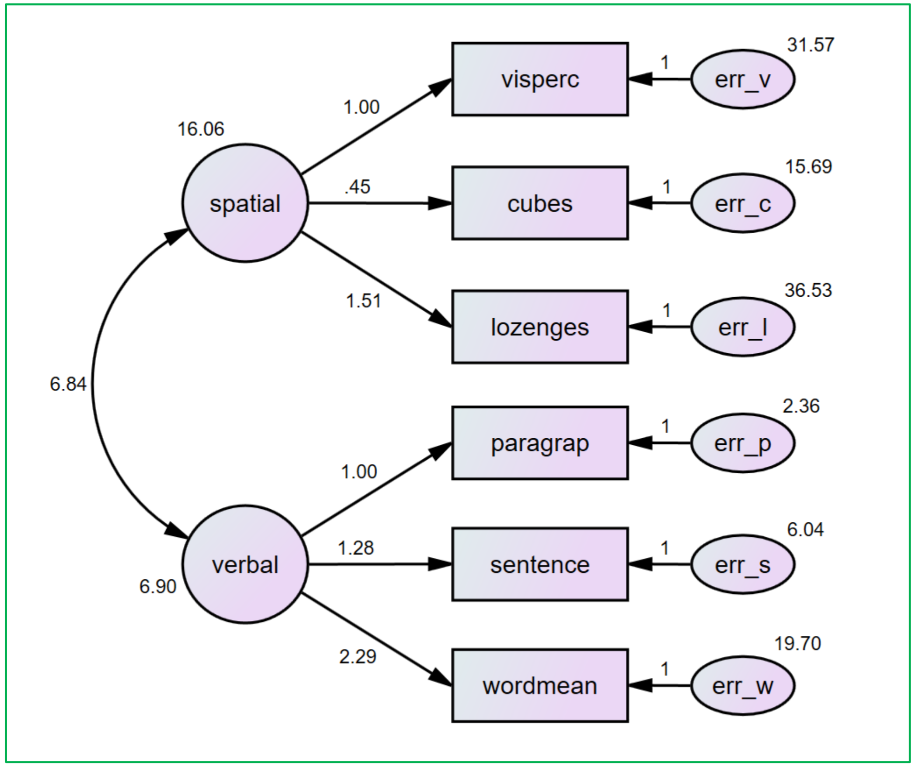


PEMEA

Step 2: Fitting the CFA Model Separately for Male and Female Groups



Unstandardized parameter estimates for **Female Group**



Unstandardized parameter estimates for **Male Group**

Fit Statistics for this Configural Invariance Model:

Chi-square = 16.480 (16 df), p = .420
CFI =.998; TLI= .997
RMSEA =.014; PCLOSE =.745

Yes, configural invariance holds!

The Metric Invariance

Metric invariance, also called factor loading invariance (or weak invariance), is demonstrated when two or more scales measure the same object using identical units (e.g., kilograms), yielding consistent readings across groups or time points.

Metric Invariance in Psychological Measurement Scale:

Metric invariance can be ensured by meeting the following conditions:

- **Configural Invariance:** Establishing configural invariance is the foundational requirement.
- **Equal Factor Loadings:** Factor loadings must be consistent across subgroups or over different time points.
- **Model Fit:** Metric invariance is confirmed if applying constraints for equal factor loadings across groups does not lead to a significant deterioration in model fit.

Steps in Testing Metric Invariance.

1. Establish Configural Invariance: This is a prerequisite. There is no need to test for metric invariance if configural invariance does not hold.

2. Assess Model Fit with Constrained Equal Factor Loadings Across Subgroups: If the model fit does not deteriorate significantly compared to the configural invariance model, then metric invariance holds. Otherwise, metric invariance does not hold.

2.1. Compute for the Chisquare Difference:

$$\Delta\chi^2 = \chi^2_{metric} - \chi^2_{configural} \qquad \Delta df = df_{metric} - df_{configural}$$

2.2. Get the p-value associated with $\Delta\chi^2$ and Δdf:

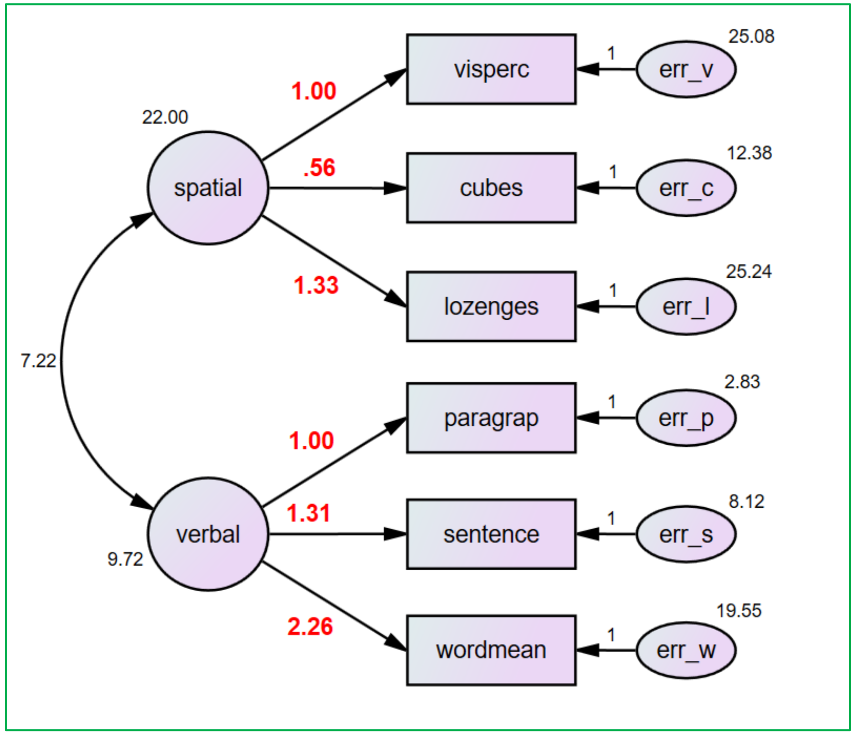
The chisquares do not differ significantly (i.e., the model fit does not deteriorate significantly) if the p-value is greater than .05. Otherwise, the model fit deteriorates significantly.

Example: Testing Metric Invariance

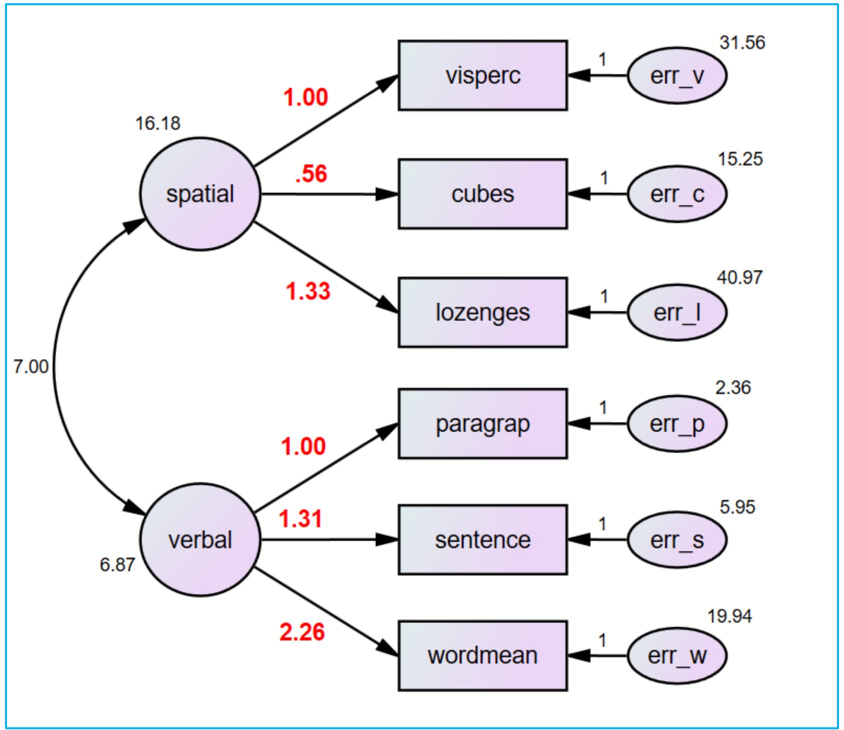
- **Step 1: Establishing Configural Invariance**

Configural invariance was already established!

Step 2: Assessing Model Fit with Constrained Equal Factor Loadings for Male & Female Groups



Unstandardized parameter estimates for **Female Group**



Unstandardized parameter estimates for **Male Group**

Testing Chisquare difference:

$\Delta\chi^2 = \chi^2_{metric} - \chi^2_{configural}$

$\Delta\chi^2 = 18.291 - 16.480$

$\Delta\chi^2 = 1.811$

$\Delta df = df_{metric} - df_{configural}$

$\Delta df = 20 - 16$

$\Delta df = 4$

P-value = .770

Yes, metric invariance holds!

Fit Statistics for this Metric Invariance Model:

Chi-square = 18.291 (20 df), p = .568
CFI =1.000; TLI= 1.008
RMSEA =.000; PCLOSE =.866

The Scalar Invariance

- Scalar invariance is also called **intercept invariance**. Intercept means starting point.

- **Scalar invariance** is demonstrated when two or more scales exhibit metric invariance and have **equal intercepts across groups.**

- **Scalar non-invariance** is demonstrated if the scales have unequal intercepts across groups.

Weighing Scale Analogy

- The primary goal is to test whether taking weight reduction pills for 10 days is effective in reducing the weight of 40-year-old mothers. The experiment involved two groups of mothers, with 20 mothers in each group. Group 1 did not take the pills, while Group 2 did. At the end of 10 days, the weights were measured using Weighing Scale 1 for Group 1 and Weighing Scale 2 for Group 2.
- Of course, the weights gathered are influenced by the calibration of the weighing scales, which may affect the accuracy and comparability of the results between the two groups.
- **Important:** The researcher cannot test the significant difference between groups because the data are flawed due to the scales having different starting points.

Weighing Scale 1 starts at 0g, which represents a properly calibrated scale.

Weighing Scale 2 starts at 70g, meaning it shows 5 kg even when there's nothing on it.

Scale Invariance in Psychological Measurement Scale:

Scale invariance can be ensured by meeting the following conditions:

- **Metric Invariance:** Establishing this serves as the foundational requirement.
- **Equal Intercepts**: Intercepts should be consistent across subgroups or at different time points.
- **Model Fit:** Scalar invariance is established if there is no significant deterioration in the model fit when constraints for equal intercepts are applied.

PEMEA

Steps in Testing Scale Invariance.

1. Establish Metric Invariance: This is a prerequisite. There is no need to test for scale invariance if metric invariance does not hold.

2. Assess Model Fit with Constrained Equal Intercepts Across Subgroups: If the model fit does not deteriorate significantly compared to the metric invariance model, then scale invariance holds. Otherwise, scale invariance does not hold.

2.1. Compute for the Chisquare Difference:

$$\Delta\chi^2 = \chi^2_{scale} - \chi^2_{metric}$$

$$\Delta df = df_{scale} - df_{metric}$$

2.2. Get the p-value associated with $\Delta\chi^2$ and Δdf:

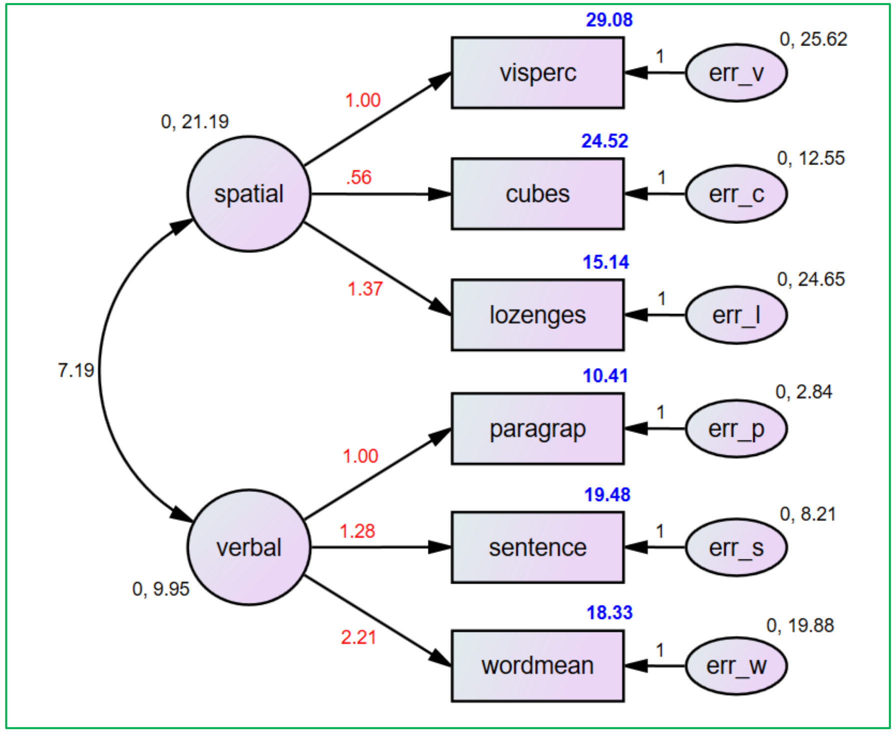
The chisquares do not differ significantly (i.e., the model fit does not deteriorate significantly) if the p-value is greater than .05. Otherwise, the model fit deteriorates significantly.

Example: Testing Scale Invariance

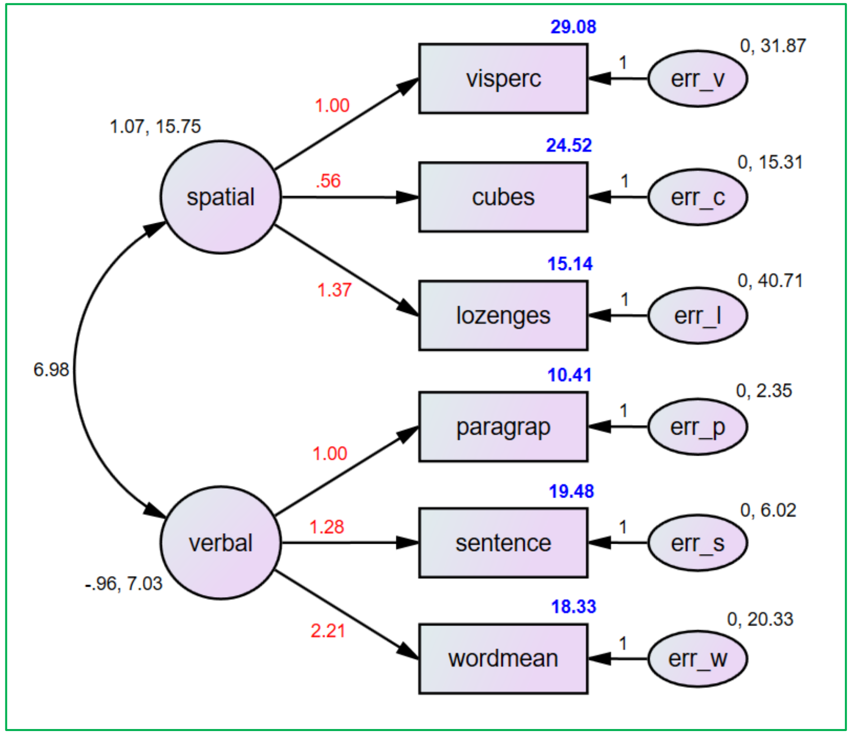
- **Step 1: Establishing Metric Invariance. This is a pre-requisite.**

Metric invariance was already established!

Step 2: Assessing Model Fit with Constrained Equal Intercepts for Male & Female Groups



Unstandardized parameter estimates for **Female Group**



Unstandardized parameter estimates for **Male Group**

Testing Chisquare difference:

$\Delta\chi^2 = \chi^2_{scale} - \chi^2_{metric}$

$\Delta\chi^2$ =22.593 - 18.291

$\Delta\chi^2$ = 4.302

$\Delta df = df_{scale} - df_{metric}$

Δdf =24 - 20

Δdf = 4

P-value = .367

Yes, scale invariance holds!

Fit Statistics for this Scale Invariance Model:

Chi-square = 22.593 (24 df), p = .544
CFI =1.000; TLI= 1.006
RMSEA =.000; PCLOSE =.873

The Strict Invariance

- **Weighing Scale Analogy:** Weighing scales must have not only the same structure and setup, the same measurement units, and the same starting point, but also the same level of precision to ensure accurate comparisons.

- **A psychological measurement scale** demonstrates strict invariance if it meets similar criteria: identical items and factor structure, equal factor loadings, equal intercepts, and equal error variances (precision) across groups or time points.

- **Strict invariance (also called uniqueness invariance)** requires meeting the characteristics of **configural invariance**, **metric invariance**, **scalar invariance**, and having equal **error variances** (precision) across groups or time points.

Strict Invariance in Psychological Measurement Scale:

Scale invariance can be ensured by meeting the following conditions:

- **Configural Invariance + Metric Invariance + Scale Invariance:** Establishing these serve as the foundational requirements.
- **Error variances** should be consistent across subgroups or at different time points.
- **Model Fit:** Strict invariance is established if there is no significant deterioration in model fit when moving from scalar invariance to a more restricted model, such as one with equal error variances as an additional constraint.

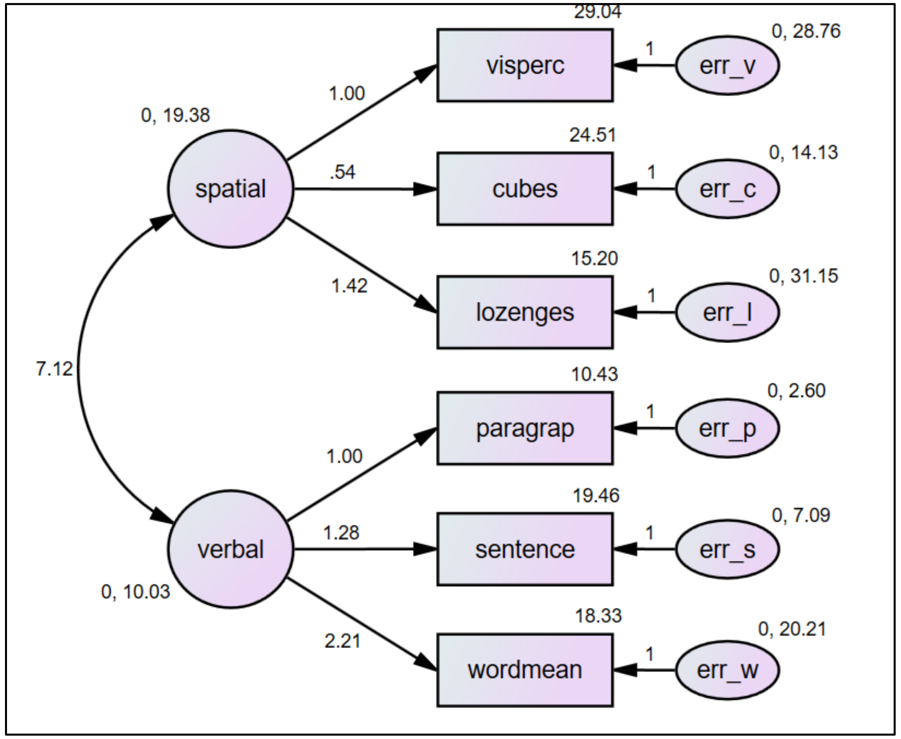
Example: Testing Strict Invariance

- **Step 1: Establishing Scale Invariance. This is a pre-requisite.**

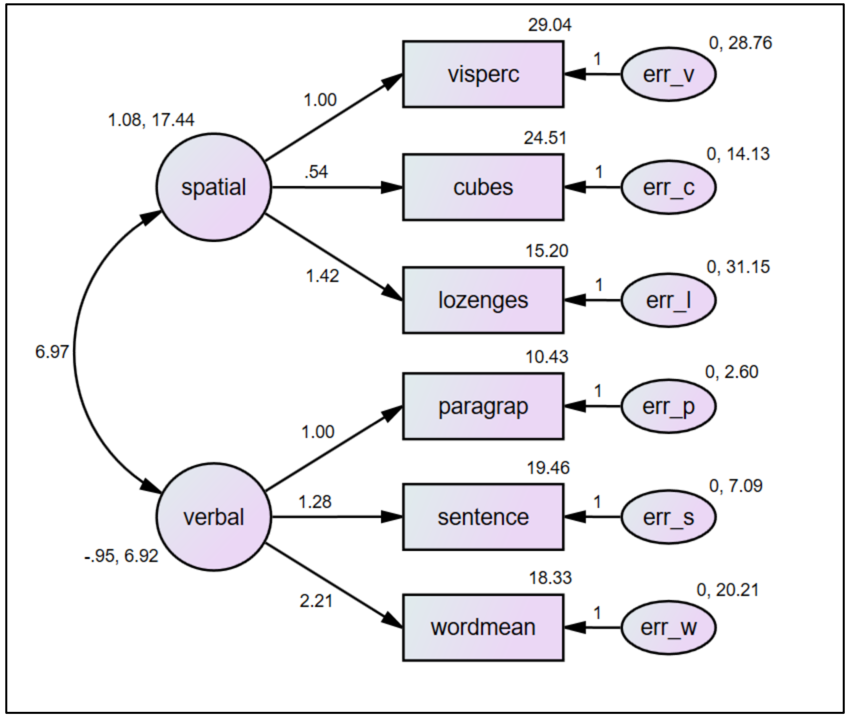
Scale invariance was already established!

Note: Scale Invariance = Configural Invariance + Metric Invariance + equal Intercepts as additional constraint.

Step 2: Assessing Model Fit with Constrained Equal Intercepts for Male & Female Groups



Unstandardized parameter estimates for **Female Group**



Unstandardized parameter estimates for **Male Group**

Testing Chisquare difference:

$\Delta\chi^2 = \chi^2_{strict} - \chi^2_{scale}$

$\Delta\chi^2$ =27.103 - 22.593

$\Delta\chi^2$ = 4.51

$\Delta df = df_{strict} - df_{scale}$

Δdf =30-24

Δdf = 6

P-value = .608

Yes, strict invariance holds!

Fit Statistics for this Scale Invariance Model:

Chi-square = 27.103 (30 df), p = .618
CFI =1.000; TLI= 1.009
RMSEA =.000; PCLOSE =.925

Summary: Practical Guide for Testing Measurement Invariance

Measurement Invariance Models	Requirements				
	CFA Model Fits Well with All Data	CFA Model Fits Well Across Subgroups' Data Separately	Equal Factor Loadings	Equal Intercepts	Equal Error Variance
Model 1: Configural Invariance	✓	✓	NA	NA	NA
Model 2: Metric Invariance	✓	✓	✓	NA	NA
Model 3: Scalar Invariance	✓	✓	✓	✓	NA
Model 4: Strict Invariance	✓	✓	✓	✓	✓

Source: Created by the author/resource person. NA = Not applicable

PEMEA

Measurement Invariance Levels (page 1 of 2)

A. Configural Invariance Levels	Description
No Configural Invariance	• The CFA model fits well for the whole group, **but fails to fit adequately across subgroups.**
Full Configural Invariance	• The CFA model fits well for the whole group. Additionally, CFA model fits well across all subgroups.
B. Metric Invariance Levels	**Description**
No Metric Invariance	• All factor loadings vary significantly across groups, meaning that the assumption of equal factor loadings does not hold. As a result, **full configural invariance** is retained.
Partial Metric Invariance	• Full invariance holds. Additionally, some factor loadings are equal across groups; others vary.
Full Metric Invariance	• Full invariance holds. Additionally, all factor loadings are equal across groups.

Source: Created by the author/resource person.

Measurement Invariance Levels (page 2 of 2)

C. Scalar Invariance Levels	Description
No Scalar Invariance	• Full metric invariance holds, but the intercepts vary across subgroups (i.e., the assumption of equal starting points does not hold). As a result, full metric invariance is retained.
Partial Scalar Invariance	• Full metric invariance holds. Additionally, some, but not all, intercepts are equal across subgroups.
Full Scalar Invariance	• Full metric invariance holds. All intercepts are equal across subgroups.
D. Strict Invariance Levels	**Description**
• No Strict Invariance	• Full scalar holds, but the error variances vary across subgroups. As a result, full scalar invariance is retained.
• Partial Strict Invariance	• Full scalar holds and not all error variances in the CFA model are equal across subgroups.
• Full Strict Invariance	• Full scalar holds and all error variances in the CFA model are equal across subgroups.

Minimum MI Requirements for Various Multiple Group Analysis (MGA) in SEM

MGA in SEM	**Minimum Measurement Invariance Requirement**
1. MGA of moderation pathways	At least Full Configural Invariance.
2. MGA of correlations between latent variables	At least Partial Metric Invariance.
3. MGA of direct pathways between latent variables	At least Partial Metric Invariance.
4. MGA of mediation pathways	At least Partial Metric Invariance.
5. MGA of latent means	At least Full Scalar Invariance.
6. MGA of sum scores or mean scores	At least Full Scalar Invariance
7. MGA of correlations between error terms	At least Partial Strict Invariance.

PEMEA

Minimum MI Requirements for One-Group Analysis in SEM

SEM with the Whole Group (Not a multiple group comparison in SEM)	**Minimum Measurement Invariance Requirement**
1. Moderation pathways	Any Measurement Invariance levels are not required.
2. Correlations between latent variables	Any Measurement Invariance levels are not required
3. Direct Effects between latent variables	Any Measurement Invariance levels are not required
4. Mediation pathways (Indirect effects)	Any Measurement Invariance levels are not required
5. Correlations between error terms	Any Measurement Invariance levels are not required

PEMEA

Testing Measurement Invariance using IBM SPSS Amos: A Demonstration

PEMEA

By: JOHNNY T. AMORA || johnny.amora@gmail.com
2024 NCEME || De La Salle University
Day3 Session: August 31, 2024

Conclusion

- We briefly reviewed **Confirmatory Factor Analysis (CFA)**. In CFA, we begin with a theory or CFA model, gather data, and use that data to test competing models, including the Single Factor Model, Correlated Factor Model, Non-Correlated Factor Model, Second-order Factor Model (Hierarchical Factor Model), and Bifactor Model.
- When testing the CFA model, factor loadings should be statistically significant ($p < .05$), and the model should demonstrate a good fit with the data (e.g., chi-square $p > .05$). Additionally, **convergent validity**, **discriminant validity**, and **reliability** should be established.

- The discussion of **Measurement Invariance (MI)** used the Correlated Factor Model with the Mental Ability Test described by Holzinger & Swineford (1939) as an example. The dataset from the **SPSS Amos User's Guide** was utilized.

- **Measurement Invariance**, as an extension of CFA, was explained using a weighing scale analogy to facilitate understanding of the concepts. MI models include Configural Invariance, Metric Invariance, Scalar Invariance, and Strict Invariance. A practical guide for testing each MI model was presented with illustrative examples prepared by the resource person, using outputs from Amos 29.

- **Measurement Invariance (MI) levels** for each MI model were discussed, including no invariance, partial invariance, and full invariance.

- The **minimum MI requirements** for simple SEM analysis and Multiple Group Analysis (MGA) in SEM were covered.

- A software demonstration using **IBM SPSS Amos 29** was conducted to test for Measurement Invariance with the **Holzinger & Swineford (1939)** dataset.

Q & A

Johnny T. Amora – *Resource Person*
Director for Institutional Effectiveness and Research
De La Salle–College of Saint Benilde, Manila
johnny.amora@benilde.edu.ph || johnny.amora@gmail.com

DE LA SALLE-
COLLEGE
OF SAINT
BENILDE

PEMEA

2024 National Conference on Educational Measurement and Evaluation
By the Philippine Educational Measurement and Evaluation Association

De La Salle University-Manila || Day 3: August 31, 2024

References

Jöreskog (1971): Jöreskog, K. G. (1971). Statistical analysis of sets of congeneric tests. Psychometrika, 36(2), 109-133.

Byrne, B. M., Shavelson, R. J., & Muthe´n, B. O. (1989). Testing for equivalence of factor covariance and mean structures: The issue of partial measurement invariance. Psychological Bulletin, 105, 456–466.

Meredith (1993): Meredith, W. (1993). Measurement invariance, factor analysis and factorial invariance. Psychometrika, 58(4), 525-543.

Millsap (2011): Millsap, R. E. (2011). Statistical approaches to measurement invariance. Routledge.

Rens van de Schoot , Peter Lugtig & Joop Hox (2012) A checklist for testing measurement invariance, European Journal of Developmental Psychology, 9:4, 486-492, DOI: 10.1080/17405629.2012.686740

Steinmetz, H. (2013). Analyzing observed composite differences across groups. *Methodology*.