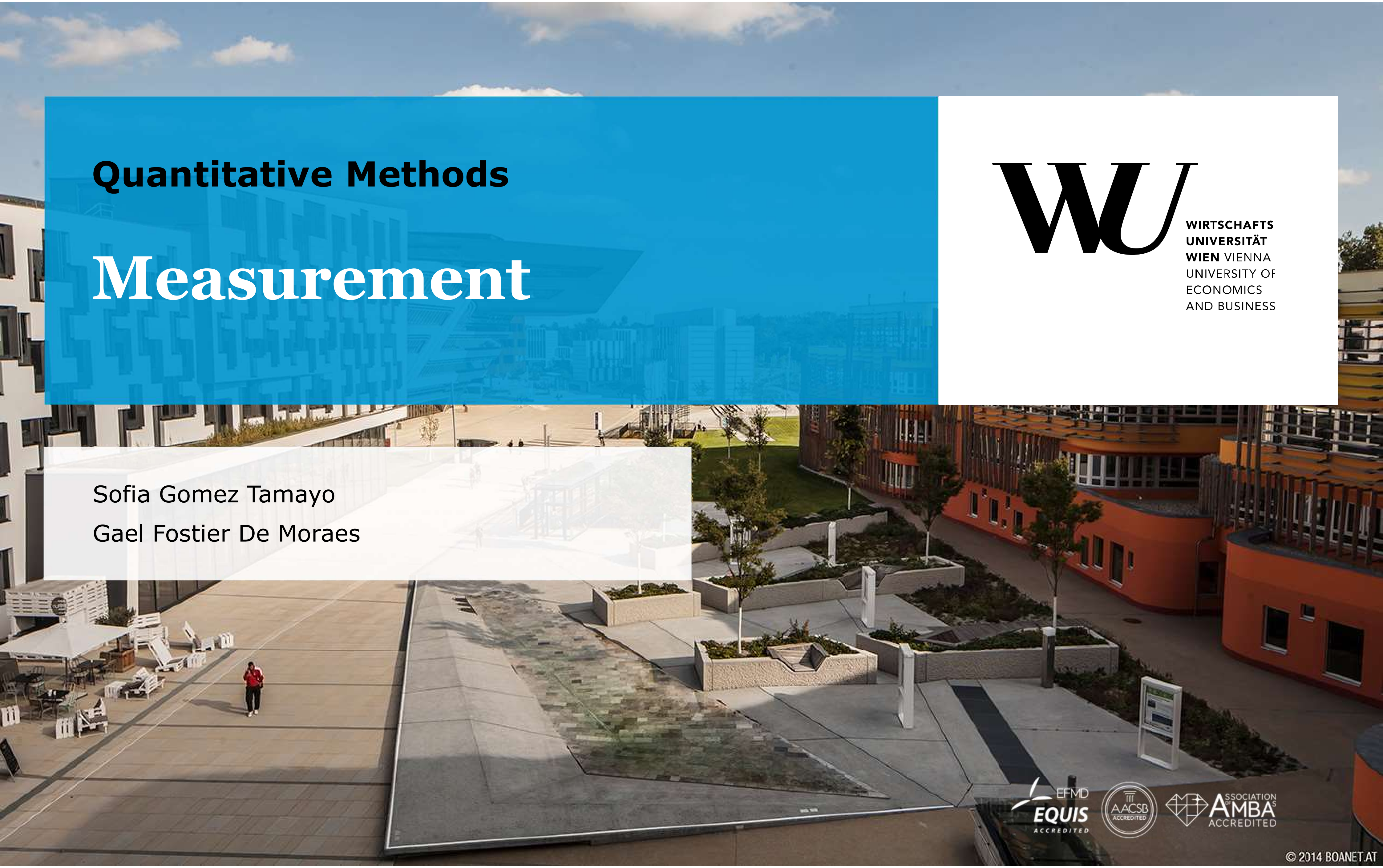


Quantitative Methods

Measurement

Sofia Gomez Tamayo
Gael Fostier De Moraes

WIRTSCHAFTS UNIVERSITÄT WIEN VIENNA UNIVERSITY OF ECONOMICS AND BUSINESS

Outline

1) **Construct validity: assessing quality of measurements**

2) **Reliability: consistency and reliability of measurement**

3) **Levels of measurement**

1) Construct validity

WIRTSCHAFTS UNIVERSITÄT WIEN VIENNA UNIVERSITY OF ECONOMICS AND BUSINESS

EFMD EQUIS ACCREDITED AACSB ACCREDITED ASSOCIATION OF MBAS AMBA ACCREDITED

WHAT IS A CONSTRUCT?

A formalization of an idea, a concept that has been defined and differentiated from other concepts

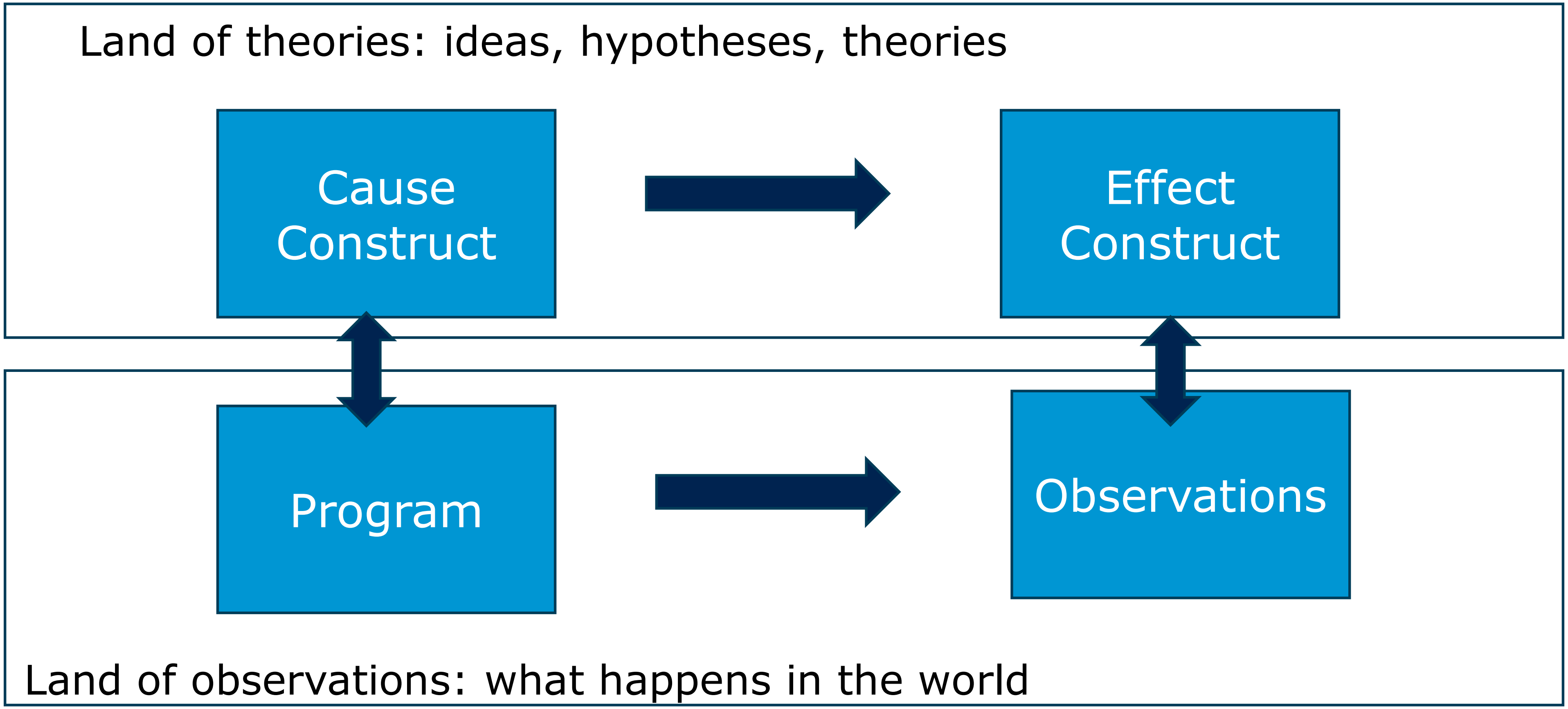
Definition of construct validity

Construct validity is the degree to which the construct translates to something real and concrete

Example:

- Construct: Unemployment
- Operationalized construct: A questionnaire to assess degrees of unemployment
- Construct validity: How much the questionnaire actually reflects the original construct of unemployment

The idea behind construct validity



Construct validity: ensuring that our theories reflect reality

Views of construct validity

Definitionalist

To have construct validity, researchers need to define the construct very precisely, distinct from all others

- Very black and white view
- Very hard to reach, e.g. self-esteem cannot relate to various concepts like confidence and worth, just one

Relationalist:

More nuanced view, things can be partly well defined, the meaning of construct changes gradually.

- Construct has to be set with a net of meaning
- Provide direct evidence that operationalization is controlled
- Provide evidence that the data supports theoretical view of the relations among constructs

Subtypes of construct validity

Translation validity: Focus on whether the operationalization reflects the construct

1) Face validity: is operationalization on its face a good translation?

- Weak option: observe a program and conclude it is a teenage pregnancy prevention program
- Stronger option: get experts to judge whether the program is designed to prevent teenage pregnancy

2) Content validity: check the operationalization against the relevant content domains for the construct

- Check the criteria that should be met to be considered a teenage pregnancy prevention program

Subtypes of construct validity

WIRTSCHAFTS UNIVERSITÄT WIEN VIENNA UNIVERSITY OF ECONOMICS AND BUSINESS

Criterion-related validity: check whether the operationalization behaves the way it should given the theory

1) **Predictive validity:** operationalization’s ability to predict something it should be able to predict (higher confidence = higher self esteem)
2) **Concurrent validity:** operationalization’s ability to distinguish between groups that it should be able to distinguish in theory
3) **Convergent validity:** operationalization’s similarity to operationalizations which should be similar (similarity of teenage pregnancy prevention program to other teenage pregnancy prevention programs)
4) **Discriminant validity:** operationalization’s similarity to operationalizations which it should differ from

Demonstrating construct validity

WIRTSCHAFTS
UNIVERSITÄT
WIEN VIENNA
UNIVERSITY OF
ECONOMICS
AND BUSINESS

EFMD EQUIS ACCREDITED AACSB ACCREDITED ASSOCIATION OF MBAS AMBA ACCREDITED

Convergent and discriminatory construct validity

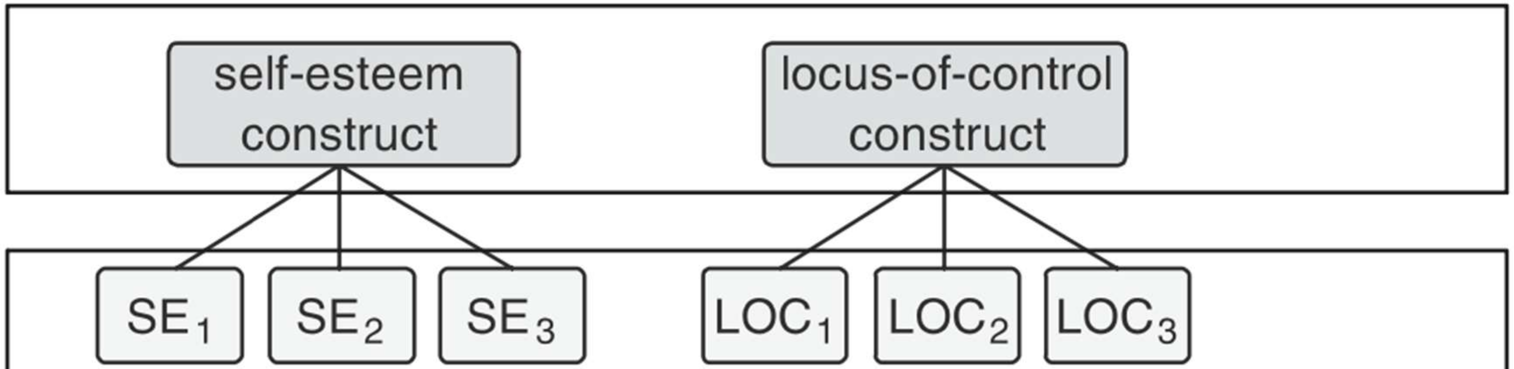
Convergent ⟷ Discriminatory

If you demonstrate one, you have demonstrated the other

To estimate the degree to which the measures are related, use correlation

- Convergence: as high as possible correlation
- Discriminant: as low as possible correlation

Correlation matrix



	SE_1	SE_2	SE_3	LOC_1	LOC_2	LOC_3
SE_1	1.00	.83	.89	.02	.12	.09
SE_2	.83	1.00	.85	.05	.11	.03
SE_3	.89	.85	1.00	.04	.00	.06
LOC_1	.02	.05	.04	1.00	.84	.93
LOC_2	.12	.11	.00	.84	1.00	.91
LOC_3	.09	.03	.06	.93	.91	1.00

The Nomological network

The network aims to link theoretical realm with the observable

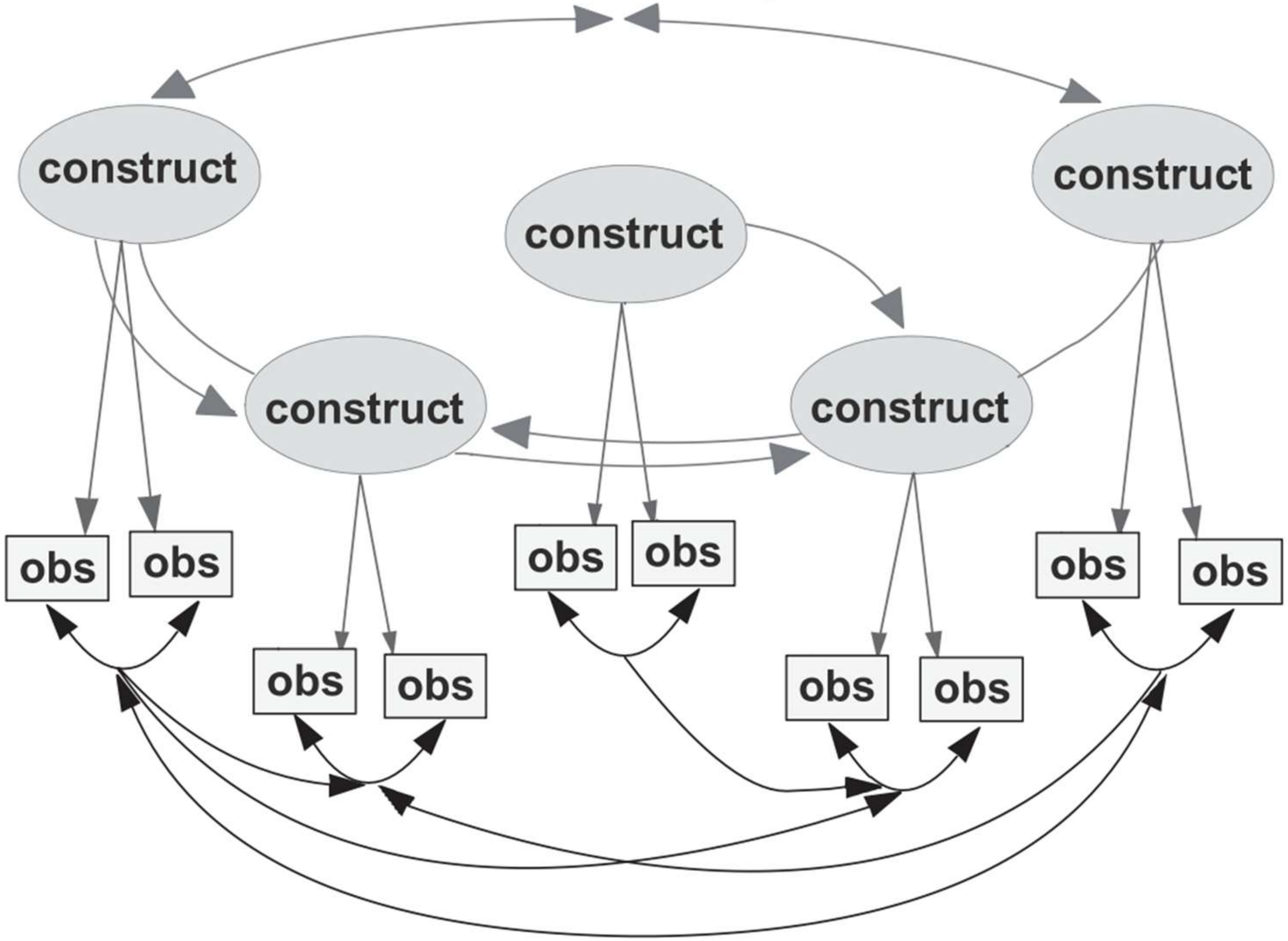
The network includes:
1. Theoretical framework for what we are trying to measure
2. Empirical framework for how to measure it
3. Specifications and linkages among and between these frameworks

Rules:
- To add a new construct or relation to a theory is that it must generate laws (nomologicals) confirmed by observables, or reduce the number of laws required
- At least some of the laws in the network must involve observables

WIRTSCHAFTS UNIVERSITÄT WIEN VIENNA UNIVERSITY OF ECONOMICS AND BUSINESS

The Nomological Network

a representation of the concepts (constructs) of interest in a study,



The Multitrait-Multimethod Matrix (MTMM)

The matrix assess both convergent and discriminant validity, to claim that measures have construct validity

The method:

- Matrix of correlations, containing the measure for each of several concepts by each of several methods
- To create the matrix, we need to arrange it by methods within concepts
- MTMM interpretation requires use of judgment, even if violating certain aspects, we can conclude strong construct validity

The Multitrait-Multimethod Matrix (MTMM)

	Traits	SE_1	SD_1	LC_1	SE_2	SD_2	LC_2	SE_3	SD_3	LC_3
		P&P			Teacher			Parent		
P&P	SE_1	(.89)								
	SD_1	.51	(.89)							
	LC_1	.38	.37	(.76)						
Teacher	SE_2	.57	.22	.09	(.93)					
	SD_2	.22	.57	.10	.68	(.94)				
	LC_2	.11	.11	.46	.59	.58	(.84)			
Parent	SE_3	.56	.22	.11	.67	.42	.33	(.94)		
	SD_3	.23	.58	.12	.43	.66	.34	.67	(.92)	
	LC_3	.11	.11	.45	.34	.32	.58	.58	.60	(.85)

The Multitrait-Multimethod Matrix (MTMM)

Components:

1. Reliability diagonals: instead of having a diagonal 1 as typically done in a matrix, we provide an estimate of the reliability of each measure

The Multitrait-Multimethod Matrix (MTMM)

Coefficients in the reliability diagonal should be the highest in the matrix

		P&P			Teacher			Parent		
	Traits	SE_1	SD_1	LC_1	SE_2	SD_2	LC_2	SE_3	SD_3	LC_3
P&P	SE_1	(.89)								
	SD_1	.51	(.89)							
	LC_1	.38	.37	(.76)						
Teacher	SE_2	.57	.22	.09	(.93)					
	SD_2	.22	.57	.10	.68	(.94)				
	LC_2	.11	.11	.46	.59	.58	(.84)			
Parent	SE_3	.56	.22	.11	.67	.42	.33	(.94)		
	SD_3	.23	.58	.12	.43	.66	.34	.67	(.92)	
	LC_3	.11	.11	.45	.34	.32	.58	.58	.60	(.85)

The Multitrait-Multimethod Matrix

Components:
1. Reliability diagonals
2. Validity diagonals: correlations between measures of the same trait measured using different methods

The Multitrait-Multimethod Matrix

The validity diagonals should be significantly different from 0 and high enough to warrant further investigation (convergence)

		P&P			Teacher			Parent		
	Traits	SE_1	SD_1	LC_1	SE_2	SD_2	LC_2	SE_3	SD_3	LC_3
P&P	SE_1	(.89)								
	SD_1	.51	(.89)							
	LC_1	.38	.37	(.76)						
Teacher	SE_2	$\boxed{.57}$	.22	.09	(.93)					
	SD_2	.22	$\boxed{.57}$	.10	.68	(.94)				
	LC_2	.11	.11	$\boxed{.46}$	.59	.58	(.84)			
Parent	SE_3	$\boxed{.56}$	.22	.11	$\boxed{.67}$	.42	.33	(.94)		
	SD_3	.23	$\boxed{.58}$	.12	.43	$\boxed{.66}$	.34	.67	(.92)	
	LC_3	.11	.11	$\boxed{.45}$	.34	.32	$\boxed{.58}$	.58	.60	(.85)

The Multitrait-Multimethod Matrix

Components:

1. Reliability diagonals
2. Validity diagonals: correlations between measures of the same trait measured using different methods
3. Heterotrait-monomethod triangles: correlations among measures that share the same method of measurement

The Multitrait-Multimethod Matrix

		P&P			Teacher			Parent		
	Traits	SE_1	SD_1	LC_1	SE_2	SD_2	LC_2	SE_3	SD_3	LC_3
P&P	SE_1	(.89)								
	SD_1	.51	(.89)							
	LC_1	.38	.37	(.76)						
Teacher	SE_2	.57	.22	.09	(.93)					
	SD_2	.22	.57	.10	.68	(.94)				
	LC_2	.11	.11	.46	.59	.58	(.84)			
Parent	SE_3	.56	.22	.11	.67	.42	.33	(.94)		
	SD_3	.23	.58	.12	.43	.66	.34	.67	(.92)	
	LC_3	.11	.11	.45	.34	.32	.58	.58	.60	(.85)

The Multitrait-Multimethod Matrix

Components:
1. Reliability diagonals
2. Validity diagonals: correlations between measures of the same trait measured using different methods
3. Heterotrait-monomethod triangles: correlations among measures that share the same method of measurement
4. Heterotrait-heteromethod triangles: correlations that differ both in trait and method

The Multitrait-Multimethod Matrix

		P&P			Teacher			Parent		
	Traits	SE_1	SD_1	LC_1	SE_2	SD_2	LC_2	SE_3	SD_3	LC_3
P&P	SE_1	(.89)								
	SD_1	.51	(.89)							
	LC_1	.38	.37	(.76)						
Teacher	SE_2	.57	.22	.09	(.93)					
	SD_2	.22	.57	.10	.68	(.94)				
	LC_2	.11	.11	.46	.59	.58	(.84)			
Parent	SE_3	.56	.22	.11	.67	.42	.33	(.94)		
	SD_3	.23	.58	.12	.43	.66	.34	.67	(.92)	
	LC_3	.11	.11	.45	.34	.32	.58	.58	.60	(.85)

The validity coefficient should be higher than the values in the columns and rows in the same triangle because trait factors are stronger than method factors

The Multitrait-Multimethod Matrix

Components:
1. Reliability diagonals
2. Validity diagonals: correlations between measures of the same trait measured using different methods
3. Heterotrait-monomethod triangles: correlations among measures that share the same method of measurement
4. Heterotrait-heteromethod triangles: correlations that differ both in trait and method
5. Monomethod blocks: all correlations with same method of measurements
6. Heteromethod blocks: all correlations that do not share a methods

Advantages and Disadvantages of MTMM

- Offers an operational methodology for assessing construct validity
- Examines how we measure in addition to what we measure

- Requires a fully crossed measurement design: each trait needs to be measured by several methods
- Two researchers might arrive at different conclusions
- It cannot quantify the degree of construct validity

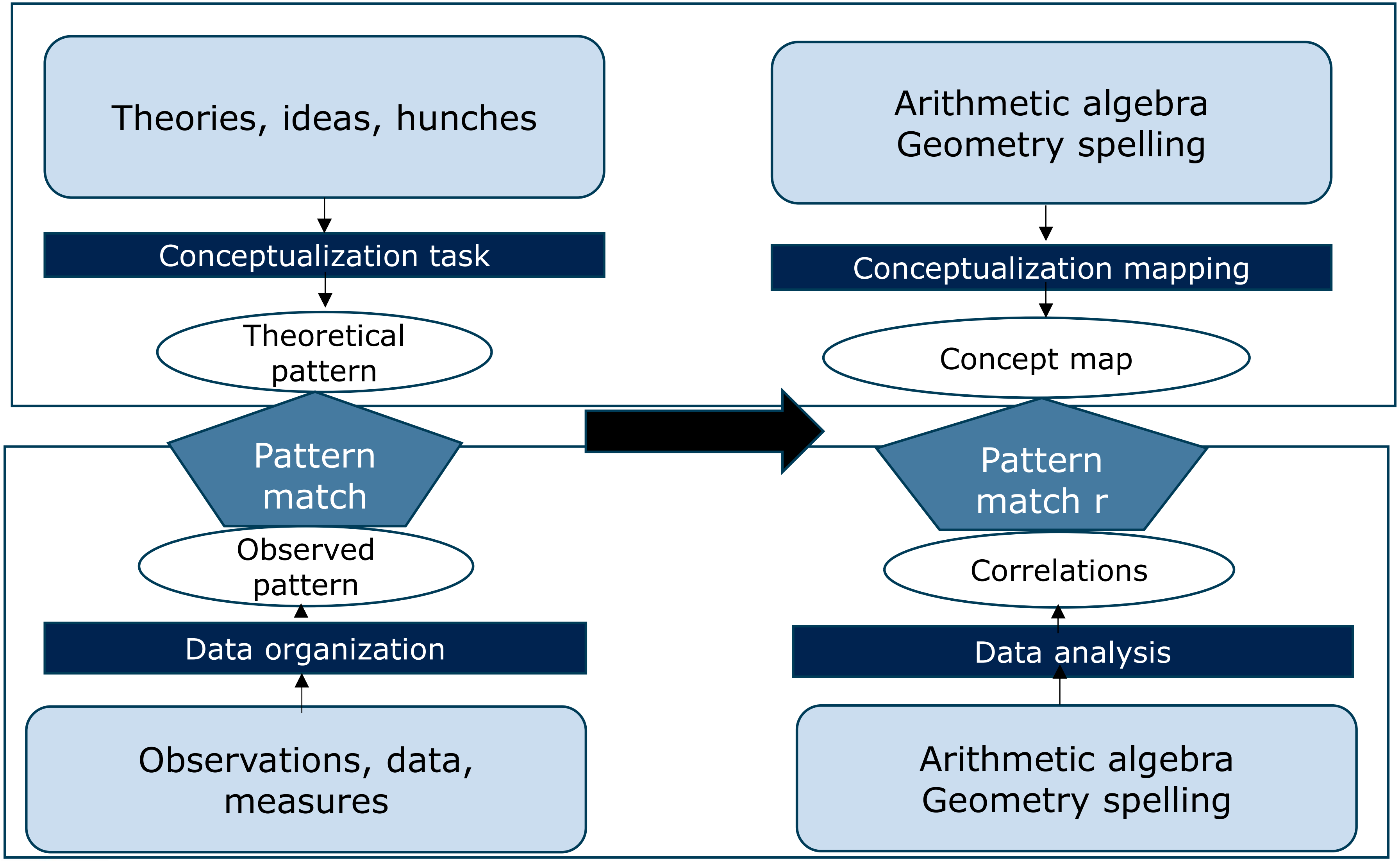
Pattern Matching for Construct Validity

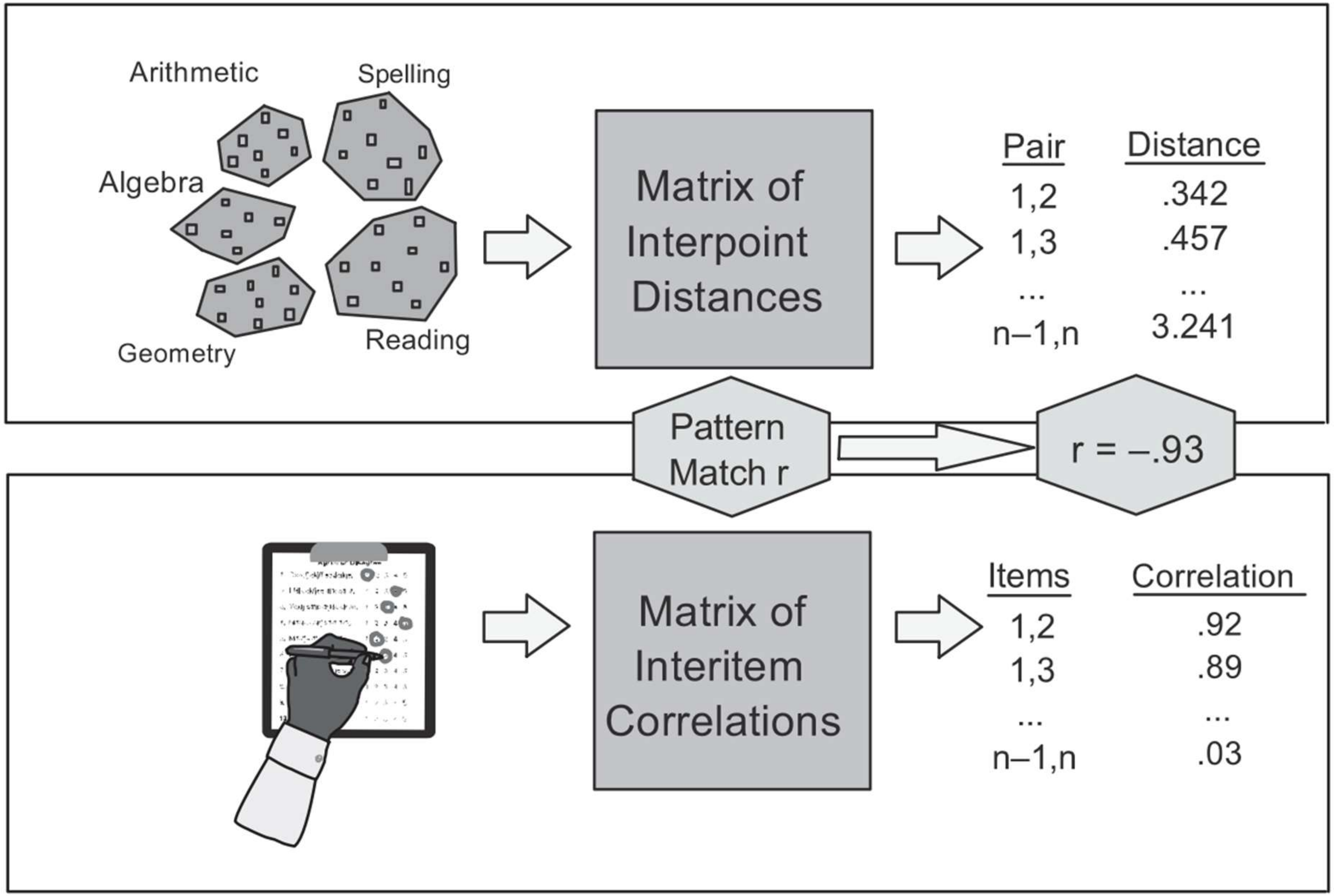
Claiming that measures have construct validity is saying that we understand how these operate in theory and that they behave as expected

→

THEY FOLLOW THE EXPECTED PATTERNS

Pattern Matching for Construct Validity




Arithmetic
Spelling
Algebra
Geometry
Reading
Matrix of Interpoint Distances
Pair
Distance
1,2
.342
1,3
.457
...
...
n–1,n
3.241
Pattern Match r
r = –.93
Matrix of Interitem Correlations
Items
Correlation
1,2
.92
1,3
.89
...
...
n–1,n
.03

Advantages and Disadvantages of Pattern Matching

- More general and flexible than MTMM
- Convergence and discrimination are continuums
- Allows to estimate the overall construct validity (the correlation of the theoretical expectations with the observed relationships)

- Requires a very well specified theory
- Needs to structure theoretical and observed patterns to observe direct correlation

Threats to construct validity

Inadequate Preoperation Explication of Constructs: Failing to define operationally well enough what was meant by the construct

Mono-operation Bias: using a single version of a program in a single place at a single point in time

Mono-method Bias: exploring only one measure of the construct

Interaction of different treatments: influence of past experiences of participants and interactions with the study

Interaction of testing and treatments: possibility of the testing or measurement influencing the reception to treatment

Threats to construct validity

Restricted generalizability across constructs: programs and measurements might affect other outcomes

Confounding constructs and levels of constructs: Lacking to specify the level of the commitment that the program involves when claiming effects

Hypothesis guessing: participants might guess what the purpose is and change their behavior

Evaluation apprehension: anxiety about evaluation and its possible role in the performance of participants

Researcher's expectancies: possible bias coming from the researcher due to their desire for certain outcomes

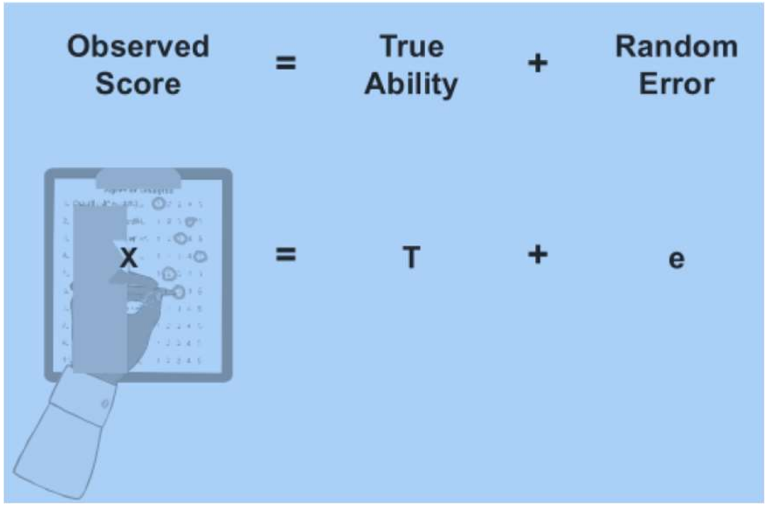
2) Reliability

Reliability

- The degree to which a measure is consistently dependable
- The degree to which it would give you the same result over and over again, assuming the underlying phenomenon is not changing

- **You cannot calculate reliability—you can only estimate it.**
- **Because of this, there are a variety of different types of reliability and multiple ways to estimate reliability for each type**

3-2a True Score Theory



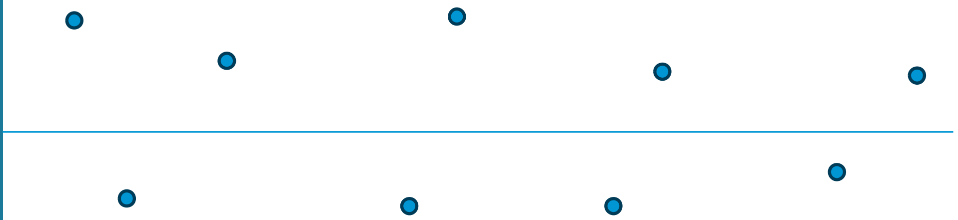
$$X = Y + e$$

- **What you observe = part explained (true ability, the true score, the legit) + part non explained**
- **Most measurement has an error component**
- **This is because what you observe has confounding factors – no measure can get 100% of the true reality**
- **Right side is non observable: only god knows**

Measurement Error

- **Good thing about random errors: they tend to null each other**
  - caused by any factors that randomly affect measurement of the variable across the sample
  - E.g. Mood swings

- **PROBLEMO: what if the errors are NOT random?**
  - Systematic errors
  - E.g., a very annoying sound during the exam
  - The more you minimize the systematic error, the closer your measurement will be from the truth



Sum = 0

In this case X = T + 0. Bingo

In reality, equation is more like:

$$\boxed{X = Y + e_r + e_s}$$

Observed = truth + random error + systematic error

Unlike random error, systematic errors tend to be either positive or negative consistently; because of this, systematic error is sometimes considered to be bias in measurement

FIGURE 3–15 Random error adds variability to a distribution but does not affect central tendency (the average)

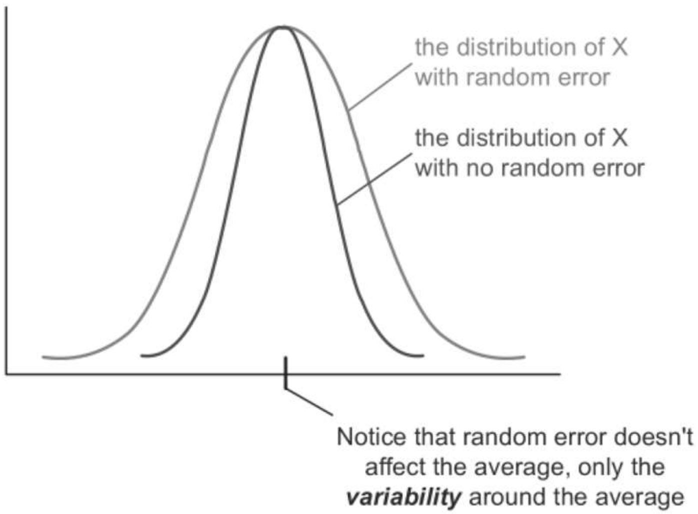
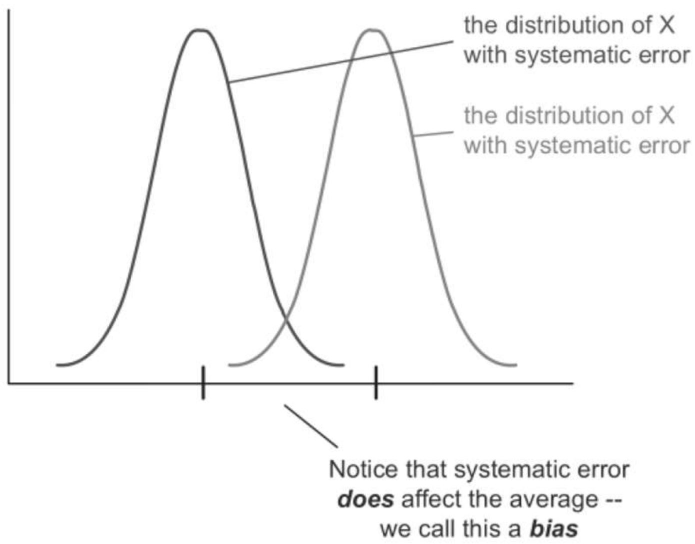


FIGURE 3–16 Systematic error affects the central tendency of a distribution



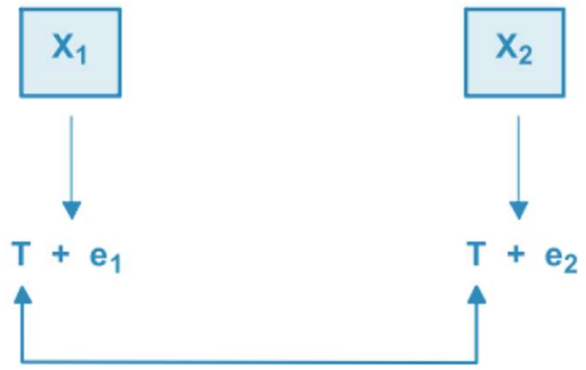
Measurement Error

- **Reducing Measurement Error**
  - **Pilot test and feedback from respondents**
  - **Interviewers and observers should be trained to not introduce bias**
  - **Double check/ double enter the data**
  - **Use statistical procedures to adjust for measurement error (simple to complex models)**
  - **Best thing: using different measures of same construct – best if different measures do not share the same systematic error**

Theory of Reliability

WU WIRTSCHAFTSUNIVERSITÄT WIEN VIENNA UNIVERSITY OF ECONOMICS AND BUSINESS

- **Repeatability or consistency.**
  - reliable if it would give you the same result over and over again (assuming that what you are measuring isn't changing).
- True score doesn't change, errors change (and that's ok) – but you should get more or less the same result every time – sometimes less, sometimes more



Reliability as the proportion of truth in your measure

$$\frac{\text{true level on the measure}}{\text{the entire measure}}$$

- But, several measures, not only one. Impossible to know
- But we can have variance!
  - Variance is a measure of the spread or distribution of a set of scores.

Equivalent to saying what's the variance of the true score, over the variance of the measure

$$\frac{\text{the variance of the true score}}{\text{the variance of the measure}}$$

Or simply: $$\frac{var(T)}{var(X)}$$

EFMD EQUIS ACCREDITED AACSB ACCREDITED AMBA ACCREDITED

Theory of Reliability

- PROBLEMO AGAIN:
  - You can't compute reliability because you can't calculate the variance of the true scores!
- DO NOT DESPAIR, scientists got our back

$$\frac{\text{covariance}(X_1, X_2)}{sd(X_1) * sd(X_2)}$$

*est: $$\frac{var(T)}{var(X)} = \frac{var(T)}{var(T) + var(e)}$$

1. You can measure X1 and you can measure X2 (consistency)
2. Covariance is nothing more than the shared variance between measures
3. We assume that the covariance of X1 and X2 is the true value, which is the only thing they have in common
4. With that the top becomes an estimation for Var (T) (yaay)
5. Variance = σ^2 = σ x σ = standard deviation x standard deviation
6. σ for X1 times σ for X2 = estimate for Var (X) (yaay 2)
7. With that we have

- If a measure is perfectly reliable, there is no error in measurement
- for a perfectly reliable measure, var (e) is zero
- True reliability = 1, no reliability = 0 (only error)
- Because of that, reliability ranges from 0 to 1

Types of Reliability

Inter-rater / inter-observer

- People are inconsistent
- How to determine whether two observers are being consistent in their observations?
- There are two major ways to actually estimate inter-rater reliability
  1. Give observers 100 data points to classify in 4 categories. Calculate the agreement rate (e.g. 86%)
  2. Two observers are observing the same phenomenon (e.g. a class) at the same intervals of time – the correlation between their rates is the reliability (consistency)
- Possibility: calibration meetings between observers to agree on scores

Test-Retest Reliability

- You administer the same test to the same (or a similar) sample on two different occasions
- Better if close in time, as things change over time.
- Correlation of results will be the reliability

Parallel-Forms Reliability

a) long list of questions, random divide in two lists, apply both to same sample, correlation is reliability

b) pre-test and post-test - Form A and Form B – randomly assign people to form A and B in pretest, switch them in post-test

Internal-Consistency Reliability

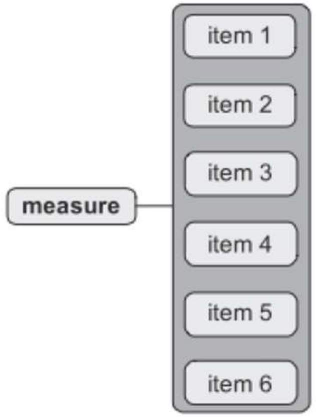
WU WIRTSCHAFTS UNIVERSITÄT WIEN VIENNA UNIVERSITY OF ECONOMICS AND BUSINESS

How consistent are the results for different items for the same construct within the measure?

❖ "item" refers to individual questions, statements, or tasks on a test or survey

- **Average Interitem Correlation**
  - You first compute the correlation between each pair of items
  - The average interitem correlation is simply the average or mean of all these correlations
- **Average Item-Total Correlation**
  - This approach also uses the interitem correlations.
  - In addition, you compute a total score for the six items and use that as a seventh variable in the analysis.

Average Inter-item Correlation

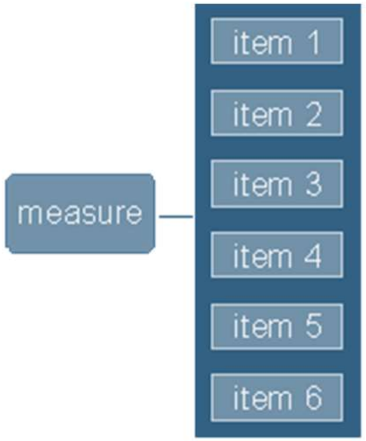


	I_1	I_2	I_3	I_4	I_5	I_6
I_1	1.00					
I_2	.89	1.00				
I_3	.91	.92	1.00			
I_4	.88	.93	.95	1.00		
I_5	.84	.86	.92	.85	1.00	
I_6	.88	.91	.95	.87	.85	1.00

.89+.91+.88+.84+.88+.92+.93+.86+.91+.95+.92+.95+.85+.87+.85=13.41

13.41 / 15 = .90

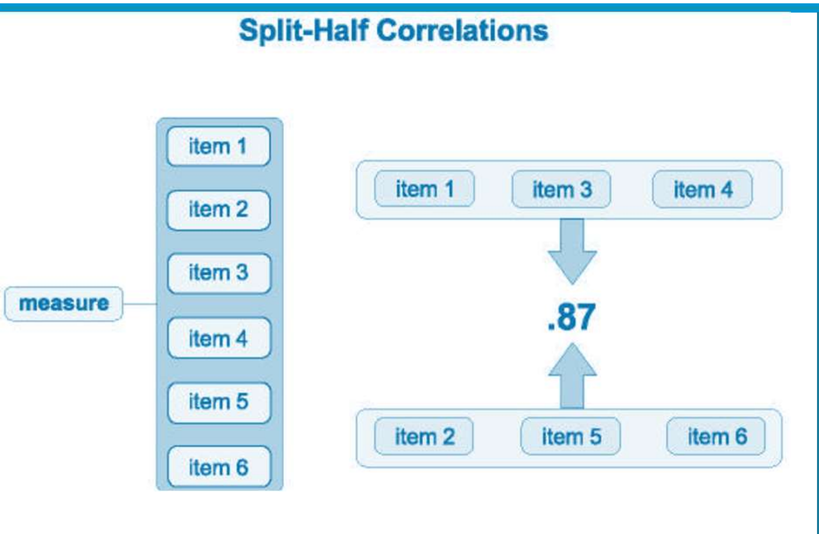
Average Item-Total Correlation



	I_1	I_2	I_3	I_4	I_5	I_6	
I_1	1.00						
I_2	.89	1.00					
I_3	.91	.92	1.00				
I_4	.88	.93	.95	1.00			
I_5	.84	.86	.92	.85	1.00		
I_6	.88	.91	.95	.87	.85	1.00	
Total	.84	.88	.86	.87	.83	.82	1.00

.85

Internal-Consistency Reliability

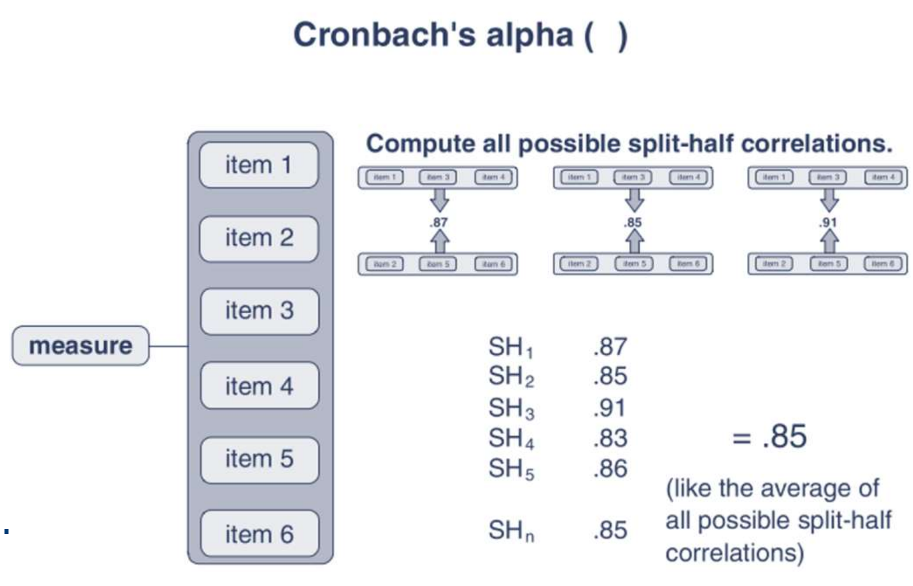


Split-Half Reliability

- You randomly divide all items that mean to measure the same construct into two sets.
- You administer the entire instrument to a sample of people and calculate the total score for each randomly divided half.
- The split-half reliability estimate, is simply the correlation between these two total scores.
- In the example, it is .87.

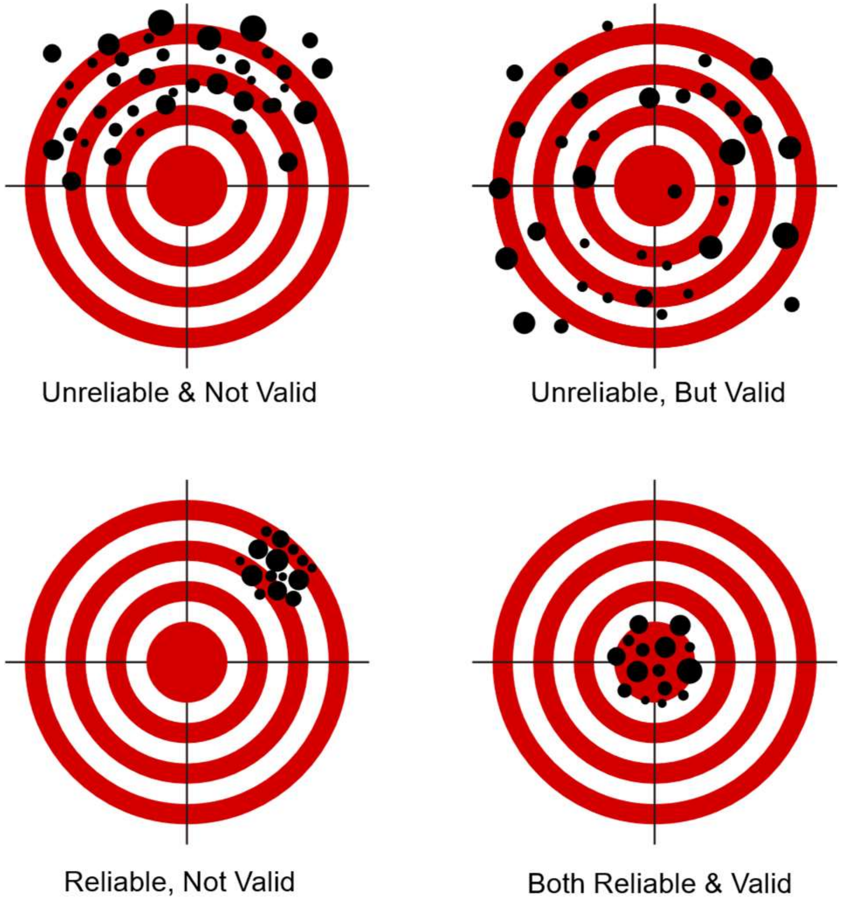
Cronbach's Alpha

- Imagine that you compute one split-half reliability and then randomly divide the items into another set of split halves and recompute
- Keep doing this until you have computed all possible split-half estimates of reliability.
- Cronbach's alpha is mathematically equivalent to the average of all possible split-half estimates
- There's a way to calculate that without going through this process



Reliability Validity

The shooting-target metaphor for reliability and validity of the measurement

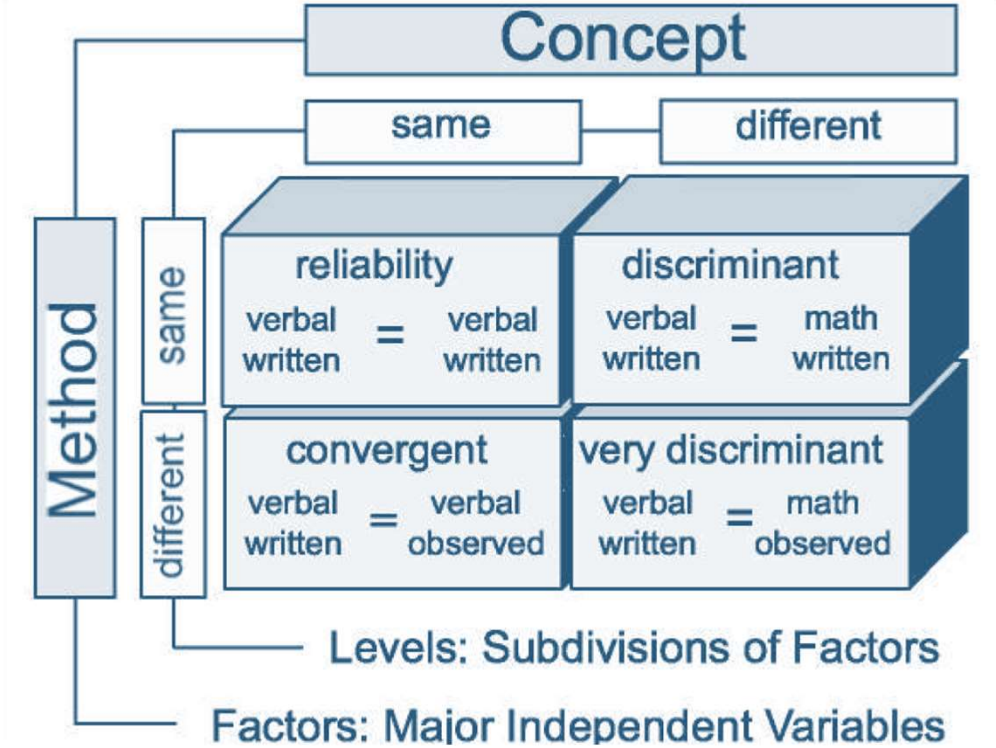


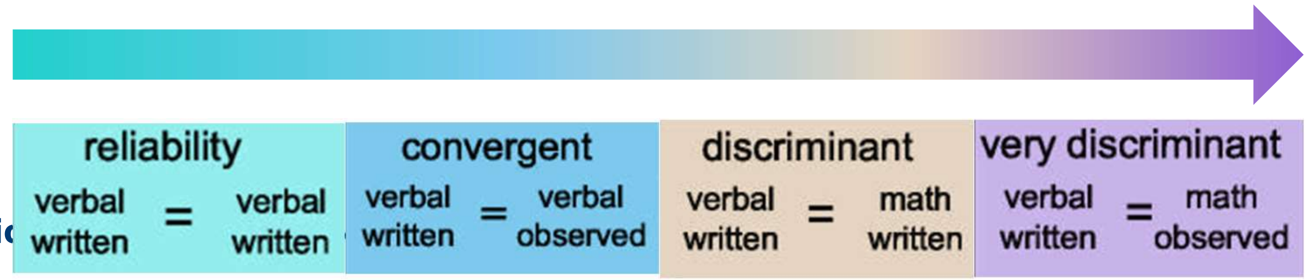
Reliability and validity are very connected

1. **you are consistently and systematically measuring the wrong value for all respondents.**
2. **you get a valid group estimate, but you are inconsistent.**
    a. **reliability is directly related to the variability of your measure.**
3. **your hits are spread across the target and you are consistently missing the center.**
4. **Your measure is both reliable and valid**

Reliability Validity

- **<u>The columns of the table</u>** indicate whether you are trying to measure the same or different concepts.
- **<u>Rows show</u>** whether you are using the same or different methods of measurement.
- **Reliability:** do two same tests find same results?
- **Convergent:** do two different methods of measurement find similar results?
- **Discriminant:** if I use same method for different concepts, results should be different
- **Very discriminant:** different methods for different concepts should be very different
- Relationships should be higher upper left until lower bottom right



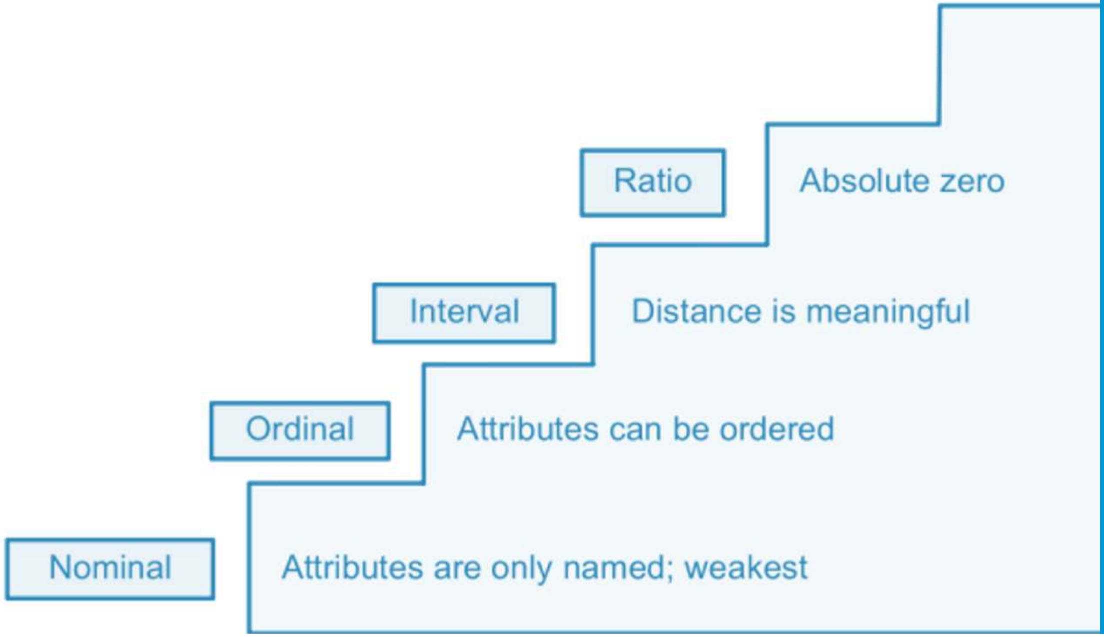


3) Levels of measurement

WIRTSCHAFTS
UNIVERSITÄT
WIEN VIENNA
UNIVERSITY OF
ECONOMICS
AND BUSINESS

Levels of Measurement

The relationship among the values that are assigned to the attributes for a variable



- **Nominal:** simply name the attribute uniquely. No ordering of the cases is implied.
  - For example, jersey numbers in basketball are measures at the nominal level.
- **Ordinal.** Attributes can be rank-ordered, but distance have no meaning.
  - E.g. Educational Attainment as 0 = less than H.S.; 1 = H.S. degree, 2 = bachelor's degree, etc.
- **Interval.** Distance DOES have a meaning and you can interpret distances.
  - When measuring temperature, the distance from 10 to 20 is same as distance from 20 to 30 degrees
- **Ratio.** Also has equal distances between measurements, but it adds a true zero point
  - means that 0 of something indicates none of it exists (like 0 kgs means no weight)
  - This allows for comparisons using multiplication or division.
  - How many clients you had in the past 6 months, are ratio variables because you can have zero clients, and it makes sense to compare numbers directly.

Summary

WIRTSCHAFTS UNIVERSITÄT WIEN VIENNA UNIVERSITY OF ECONOMICS AND BUSINESS

Construct validity: what are you really measuring? getting it right

1. **What It Means**: It's about making sure your test actually measures what you think it does.
2. **Two Parts**:
    a) **Translation validity (translating well)**: Did you turn your idea into a question or test accurately?
    b) **Predicting Things Right**: Does your test predict other outcomes as expected?

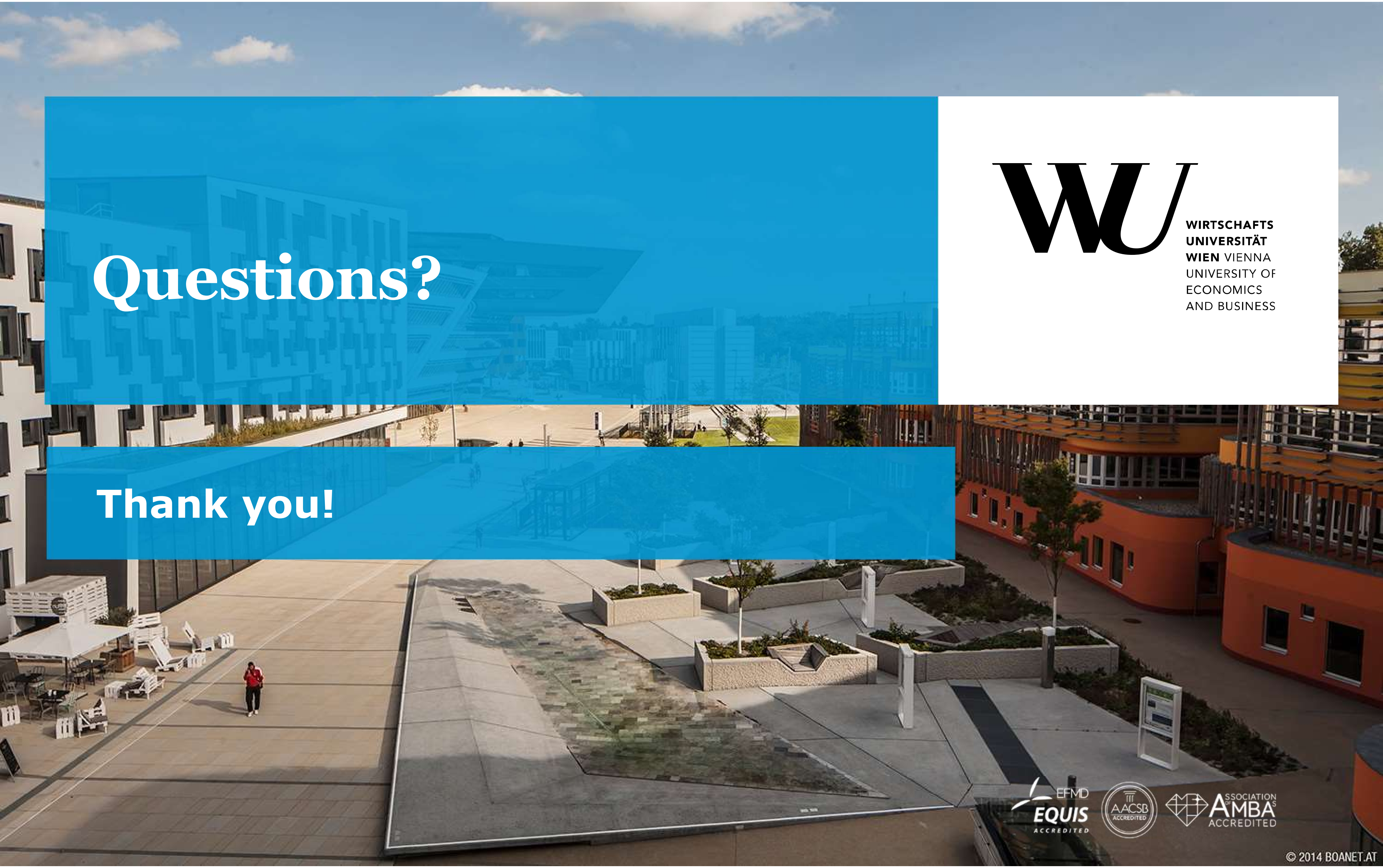
Reliability (consistency): Can You Trust Your Measurement?

1. **Basic Idea**: Every measurement is a mix of the real score and some mistakes (errors, nature).
2. **Defined As**: How much of what you measure is the real deal versus error.

Measurement Levels: Knowing What Your Numbers Mean

1. **Why It Matters**: The kind of measurement you use tells you what math you can do with it.
2. **Effect**: It decides the best way to handle and make sense of your data

Final Thought: Picking the right way to measure and understanding your results are key to good research.



Questions?

WIRTSCHAFTS UNIVERSITÄT WIEN VIENNA UNIVERSITY OF ECONOMICS AND BUSINESS

Thank you!