SCIENTIFIC REASONING: MEASUREMENT MODELS AND VALIDATION PROCESS

Rossella Caliciuri

Lehrstuhlmeeting - July 1, 2025

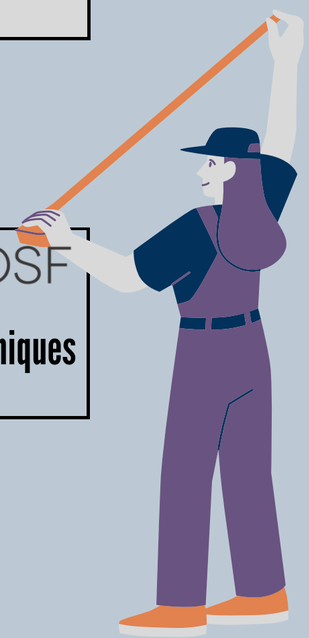
SCIENTIFIC REASONING: MEASUREMENT MODELS AND VALIDATION PROCESS

Scientific Reasoning:
An Umbrella Review of Definitions, Measurement Instruments, Psychometric Models, and Validation Practices

Scientific Reasoning Scale in Italy:
Validation Studies

under review in *Thinking Skills and Creativity*

Scientific Reasoning:
Comparison between Measurement Models and Data Analysis Techniques



BACKGROUND & RATIONALE

What is scientific reasoning?

Despite its importance, there remains a **lack of consensual understanding and different definitions**.

BACKGROUND & RATIONALE

Foundations of the Scientific Reasoning Scale (SRS)

The SRS is a is a tool created to measure if people can really **think like scientists**, not just know facts.

BACKGROUND & RATIONALE

Foundations of the Scientific Reasoning Scale (SRS)

The SRS is a is a tool created to measure if people can really **think like scientists**, not just know facts.
It uses three main frameworks:

Philosophy & Methodology of Science
How Science Should Work

It's not enough to know how to do an experiment, you also need to understand when an experiment has problems or when the results are not clear. The SRS wants to see if you are good at "spotting" weaknesses in scientific research.

Public Understanding of Science
What People Already Know About Science

There are already some tests that check if people know basic science facts or if they understand how the scientific method works. The SRS starts from this point, but wants to go further, adding more critical skills that the old tests didn't measure.

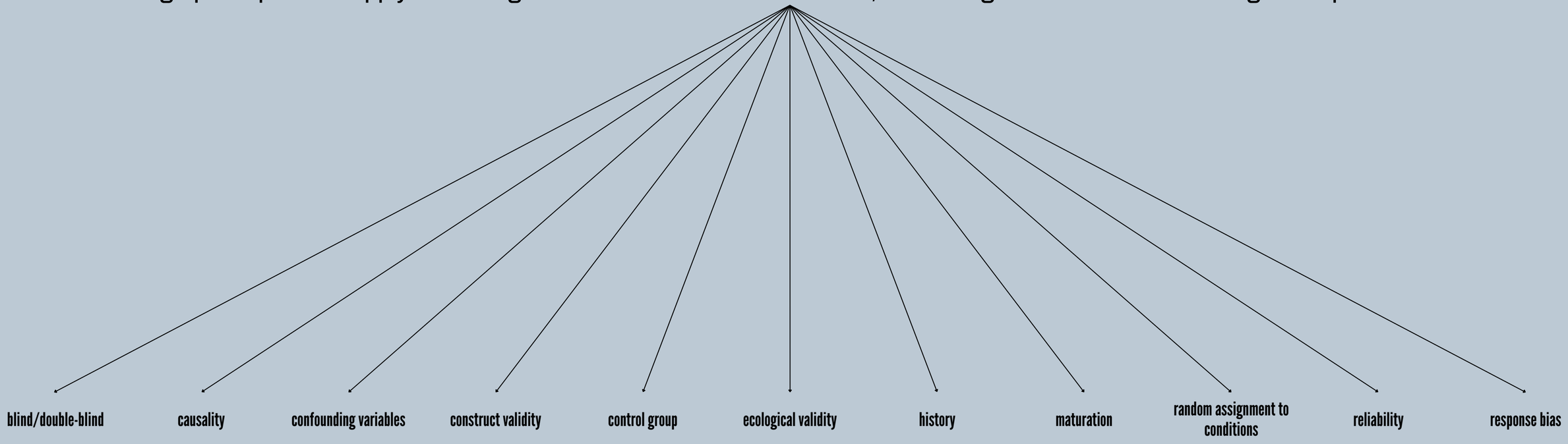
Cognitive Developmental Psychology
How People Learn Scientific Thinking

Psychologists study how people develop the ability to think like scientists – that means making hypotheses, collecting evidence, and making conclusions. The SRS tries to see if you can use these "scientific investigator" skills when you face new information.

The SRS doesn't just ask you "what you know" about science, but **"how you think" when you deal with science**.
The goal is to see if you can judge the quality of scientific information.

BACKGROUND & RATIONALE

The **Scientific Reasoning Scale** (SRS; Drummond & Fischhoff, 2017) has been originally validated in the US and then in Turkey (Muslu Kaygisiz et al., 2018). The SRS assesses **individuals' ability to evaluate scientific evidence**. The scale consists of **11 true/false items** that challenge participants to apply reasoning skills to brief scientific scenarios, evaluating core scientific reasoning concepts:



1. Double blind

In a taste test, a researcher puts Brand A coffee in a cup with white tape on it and Brand B coffee in an identical cup with black tape on it. A lab assistant gives tasters one of the cups, while the researcher watches their facial expressions.
True or False? The lab assistant should not watch the cups being filled.

2. Causality

A researcher finds that American states with larger parks have fewer endangered species.
True or **False**? These data show that increasing the size of American state parks will reduce the number of endangered species.

3. Confounding variables

A researcher has subjects put together a jigsaw puzzle either in a cold room with a loud radio or in a warm room with no radio. Subjects solve the puzzle more quickly in the warm room with no radio.
True or False? The scientist cannot tell if the radio caused subjects to solve the puzzle more slowly.

4. Construct validity

An education researcher wants to measure the general math ability of a sample of high-performing math students. All the students have taken classes in geometry and pre-calculus.
True or **False**? The education researcher can measure general math ability by giving the students a geometry test.

5. Control Group

Two scientists test an anti-acne cream on teenagers with acne. Scientist A wants to give the cream to all the teenagers in the study. Scientist B wants to give the cream to half the teenagers and give a cream without anti-acne ingredients to the other half.
True or **False**? Both ways of testing the cream are equally good.

6. Ecological validity

A researcher has a group of subjects play a competitive game. Each subject's goal is to make money by buying and selling tokens. Subjects are paid a flat fee for participating in the experiment.
True or **False**? The researcher can confidently state that the behavior in the experiment reflects real-life buying and selling behavior.

7. History

A randomly selected sample of Americans is surveyed about disease A before and after a 6-month media campaign about the disease. Mid-way through the media campaign, a famous celebrity dies of Disease A. The survey data indicate that knowledge of Disease A is higher after the campaign.
True or False? The media campaign may not have increased knowledge of Disease A.

8. Maturation

Subjects in an experiment must press a button whenever a blue dot flashes on their computer screen. At first, the task is easy for subjects. But as they continue to perform the task, they make more and more errors.
True or **False**? The blue dot must flash more quickly as the task progresses.

9. Random assignment to condition

esearchers want to see whether a health intervention helps school children to lose weight. School children are sorted into either an intervention or control group.
True or **False**? The researchers should assign the overweight children to the intervention group.

10. Reliability

A researcher develops a new method for measuring the surface tension of liquids. This method is more consistent than the old method.
True or **False**? The new method must also be more accurate than the old method.

11. Response bias

Two researchers are developing a survey to measure consumers' feelings about customer service. Researcher A wants customers to rate their agreement with the statement "I am satisfied with customer service" on a 5-point scale, where 1 = strongly agree and 5 = strongly disagree. Researcher B wants customers to rate customer service on a 5-point scale, where 1 = not dissatisfied at all and 5 = highly dissatisfied.
True or **False**? These questions are equally good for measuring how consumers feel about customer service

BACKGROUND & RATIONALE

Despite its usefulness, the SRS has yet to be validated in Italy, where **no valideted measure of scientific reasoning currently exists**.

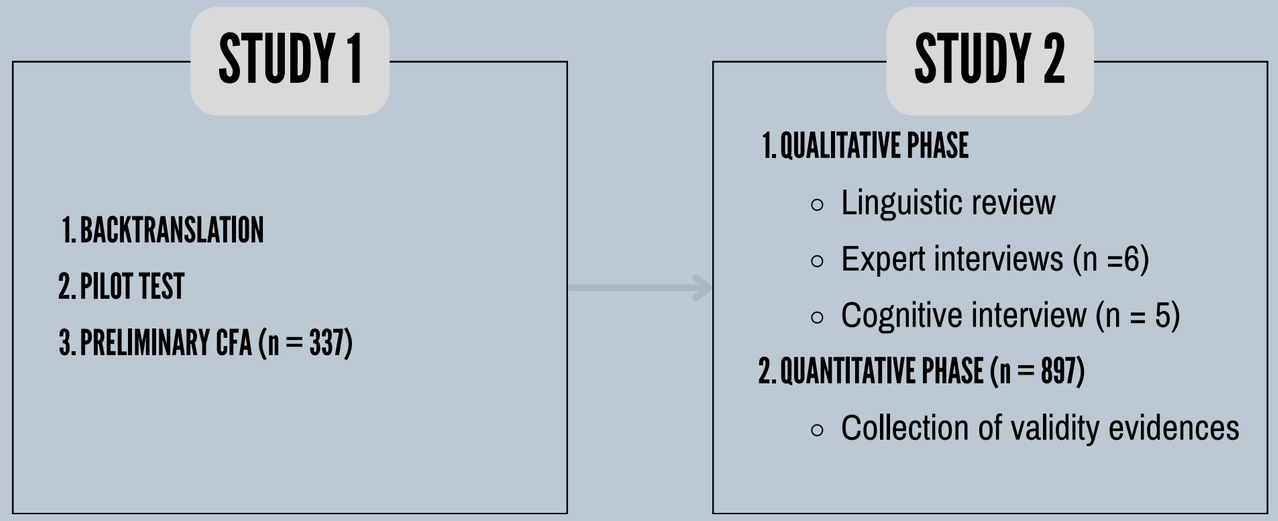
Having validated measures is essential for several reasons:

- It facilitates meaningful **comparisons across countries**, reflecting how well individuals are prepared to engage with scientific concepts and issues. This aligns with the objectives of the PISA 2025 assessment (Programme for International Student Assessment), organized by the OECD (Organisation for Economic Co-operation and Development);

- From a psychometric perspective and in line with contemporary view of validity, validating a measurement scale in diverse cultural contexts provides **additional evidence of its validity** (Hubley & Zumbo, 2011).

AIM & METHODS

To address this gap, the present study aims to **validate the Italian version of the SRS**.
A **multi-study**, **multi-method** approach was employed:



AIM & METHODS

To address this gap, the present study aims to **validate the Italian version of the SRS**.
A **multi-study**, **multi-method** approach was employed:

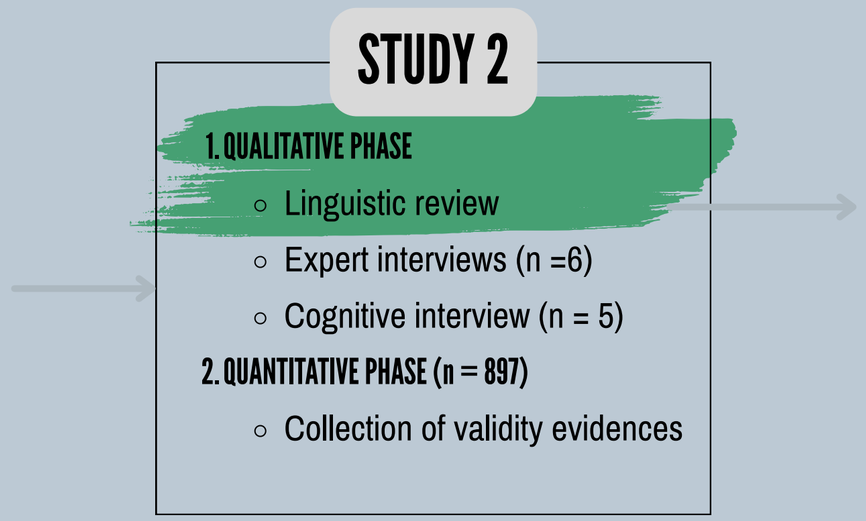
STUDY 1

1. BACKTRANSLATION
2. PILOT TEST
3. PRELIMINARY CFA (n = 337)

We assessed the scale's factor structure and identified **several psychometric weaknesses**: some items did not achieve the required factor loading cut-off points.

AIM & METHODS

To address this gap, the present study aims to **validate the Italian version of the SRS**.
A **multi-study**, **multi-method** approach was employed:



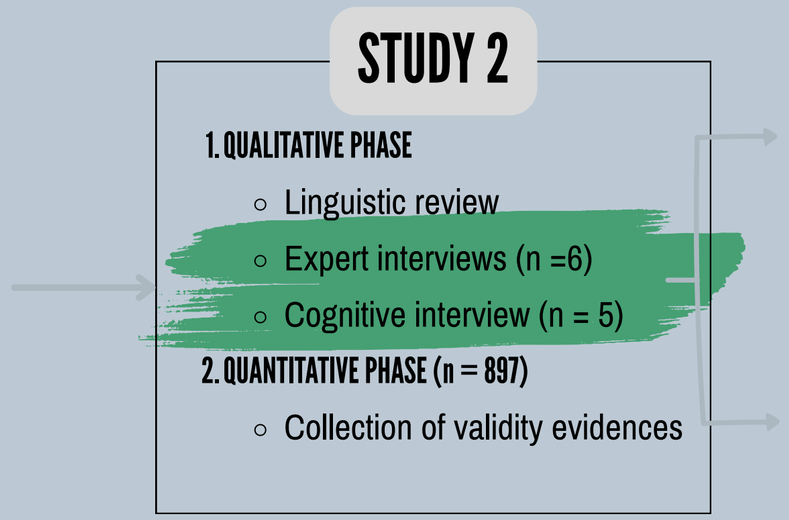
PREVIOUS VERSION: *In a taste test, a researcher puts Brand A coffee in a cup with white tape on it and Brand B coffee in an identical cup with black tape on it. A lab assistant gives tasters one of the cups, while the researcher watches their facial expressions.* ***True*** *or False? The lab assistant should not watch the cups being filled.*

NEW VERSION: *In an experiment to evaluate coffee taste preferences, a researcher puts brand A coffee and brand B coffee in identical cups, placing identifying labels on the bottom to distinguish the coffee types. A lab assistant then distributes one of the cups to the tasters, while the researcher observes their facial expressions.*

- **A.** ***The lab assistant should not look at the cups while they are being filled.***
- **B.** *The lab assistant should look at the cups while they are being filled.*

AIM & METHODS

To address this gap, the present study aims to **validate the Italian version of the SRS**.
A **multi-study**, **multi-method** approach was employed:

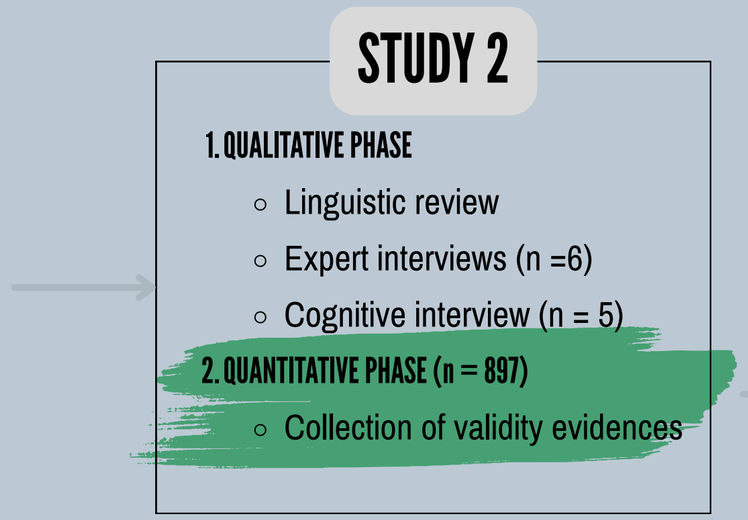


EXPERT INTERVIEWS: to gain further **insight into the construct** within the specific Italian context. Experts included full professors of research methodology in psychology, researchers in physics and statistics, and middle school teachers of mathematics and science (in line with the principle of maximum variation sampling; Patton, 2005).

COGNITIVE INTERVIEWS: to test the **comprehensibility and applicability** of the first draft of the scale to the lives of Italian adults, interviews were conducted with individuals of different genders, age ranges, and educational backgrounds from the target population. Cognitive interviews are structured interviews that focus on the cognitive processes respondents use to answer survey questions. The goal is to thoroughly explore respondents' understanding of the items and their perceptions of each item's relevance to their lives (Willis, 2004).

AIM & METHODS

To address this gap, the present study aims to **validate the Italian version of the SRS**.
A **multi-study**, **multi-method** approach was employed:



QUANTITATIVE PHASE: to assessed the psychometric properties of the scales using Classical Test Theory (**CTT**) and provides evidence of their validity through structural equation modeling (**SEM**), in line with **contemporary view of validity** (Hubley & Zumbo, 2011).

AIM & METHODS

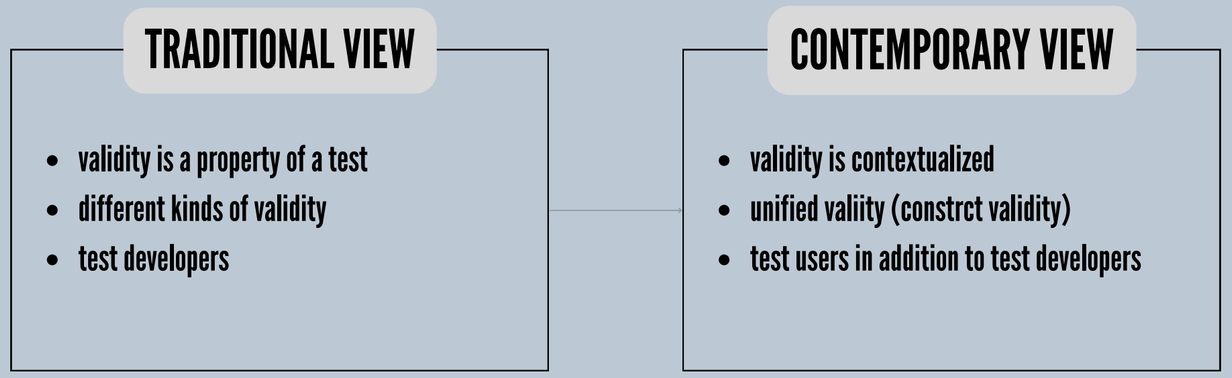
To address this gap, the present study aims to **validate the Italian version of the SRS**.
A **multi-study**, **multi-method** approach was employed:

STUDY 2

1. QUALITATIVE PHASE
   - Linguistic review
   - Expert interviews (n =6)
   - Cognitive interview (n = 5)
2. QUANTITATIVE PHASE (n = 897)
   - Collection of validity evidences

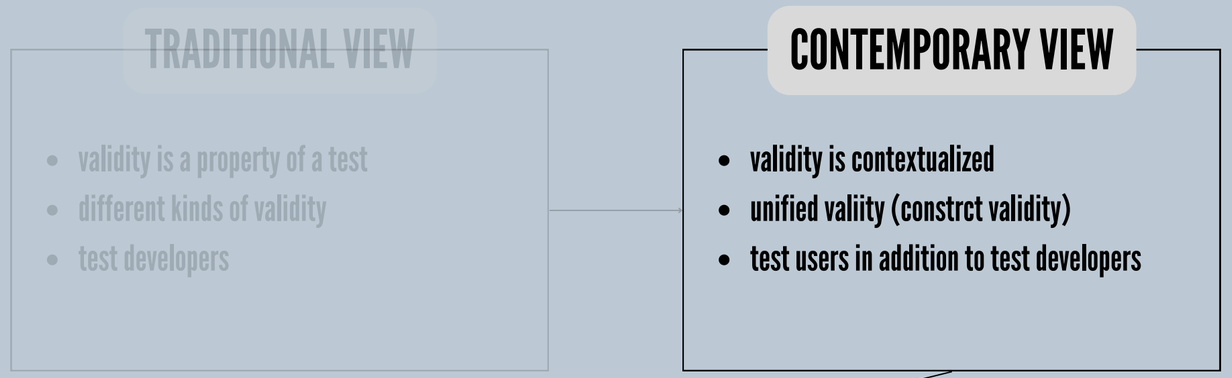
QUANTITATIVE PHASE: to assessed the psychometric properties of the scales using Classical Test Theory (CTT) and provides evidence of their validity through structural equation modeling (**SEM**), in line with **contemporary view of validity** (Hubley & Zumbo, 2011).

AIM & METHODS - CONTEMPORARY VIEW OF VALIDITY



(Cronbach & Meehl, 1955; Hubley & Zumbo, 2011; Messick, 1989; Sorgente & Zumbo, 2025; Zumbo, 2005).

AIM & METHODS - CONTEMPORARY VIEW OF VALIDITY



(Cronbach & Meehl, 1955; Hubley & Zumbo, 2011; Messick, 1989; Sorgente & Zumbo, 2025; Zumbo, 2005).

In this validation process, Structural Equation Modeling (**SEM**) is pivotal as they integrate regression, path analysis, and latent variable models.
This validation process aims to establish the validity of the Italian SRS across **various sources of evidence**:

FACTOR STRUCTURE	RELIABILITY	GENERALIZABILITY	KNOWN-GROUP	CRITERION VALIDITY	CONVERGENT VALIDITY

AIM & METHODS - SAMPLE

This sample is **representative of the Italian population aged 18 to 60** years in terms of gender distribution, age, and residency, according to the estimated ISTAT data for 2024 (Istat, Istituto Nazionale di Statistica, 2024).

AIM & METHODS - SAMPLE

This sample is **representative of the Italian population aged 18 to 60** years in terms of gender distribution, age, and residency, according to the estimated ISTAT data for 2024 (Istat, Istituto Nazionale di Statistica, 2024).

n = 897

- **GENDER**:
  - **Female: 50.5%**
  - Male: 48.94%
  - Non-binary or prefer not to disclose: 0.56%

- **AGE GROUPS**:
  - 18 – 27 (genZ): 20.18%
  - 28 – 44 (genY): 35.12%
  - **45 – 60 (genX): 44.70%**

- **REGION OF RESIDENCE**:
  - **Northwest: 25.7%**
  - Northeast: 19.6%
  - Central: 20.6%
  - South: 22.9%
  - Islands: 11.3%

- **EDUCATIONAL LEVEL**:
  - Middle School: 7.28%
  - **High School Diploma: 52.86%**
  - Bachelor's/Master's Degree: 36.28%
  - Doctoral Degree: 3.58%

- **EMPLOYMENT STATUS**:
  - High School Students: 1.37%
  - University Students: 14.06%
  - **Employed: 67.20%**
  - Unemployed: 17.37%

- ***DO YOU FEEL YOU BELONG TO A RELIGION?***
  - ***Yes:* 52.8%**
  - *No:* 47.2%

- **POLITICAL ORIENTATION**
  - **Left: 37.1%**
  - Centre: 32.9%
  - Right: 30%

AIM & METHODS - INSTRUMENT

- Survey in Qualtrics, 45 minutes
- from December 2024 to January 2025
- quality and attention control checks
- participants voluntary and compensated

SRS - italian version, 11 items

CONVERGENT

- Cognitive Reflection Test-Long (CRT-Long; Frederick, 2005; Primi et al., 2016) - 6 items
- Probabilistic Reasoning Scale (PRS; Primi, Morsanyi, Galli, & Chiesi, 2017) - 14 items

CRITERION

- Paranormal Health Beliefs Scale (PHBS, Donizzetti & Petrillo, 2017) - 31 items
  - Religious
  - Superstitious
  - Extraordinary Events
  - Parapsychological
  - Pseudo-scientific
- Climate Change Beliefs - 2 items
- Climate Change Awareness (Robba et al., 2024) - 6 items

KNOW-GROUP

- Socio-demo: gender, age, region of residence, education level, and employment status
- Political orientation - 1 item
- Religious affiliation - 1 item

All the scales have been validated or previously used in the Italian context. Additionally we verified the factorial structure in our sample by performing a CFA. **All scales showed good fit indices.** We also assessed the reliability of the scales and subscales using the composite reliability (ω). **All reliability scores range from good to excellent**, with values between 0.73 and 0.95.

AIM & METHODS - DATA ANALYSIS

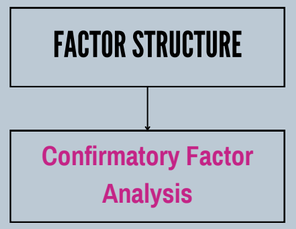
DESCRIPTIVE STATISTICS using SPSS, to summarize the characteristics of the participants.

DIFFERENT KINDS OF VALIDITY EVIDENCES through SEM. All models were performed using Mplus. ML, MLR or WLSMV were used as the estimation method, depending on whether the variables were normally distributed, non-normally distributed, or dichotomous, respectively. FIML was used as the method to handle missing data.

AIM & METHODS - DATA ANALYSIS

DESCRIPTIVE STATISTICS using SPSS, to summarize the characteristics of the participants.

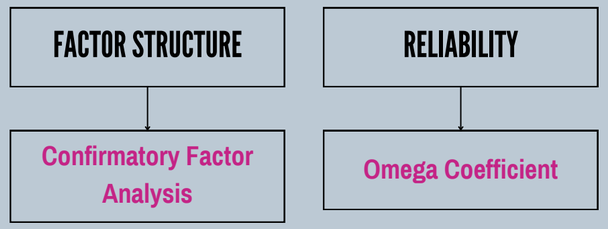
DIFFERENT KINDS OF VALIDITY EVIDENCES through SEM. All models were performed using Mplus. ML, MLR or WLSMV were used as the estimation method, depending on whether the variables were normally distributed, non-normally distributed, or dichotomous, respectively. FIML was used as the method to handle missing data.



AIM & METHODS - DATA ANALYSIS

DESCRIPTIVE STATISTICS using SPSS, to summarize the characteristics of the participants.

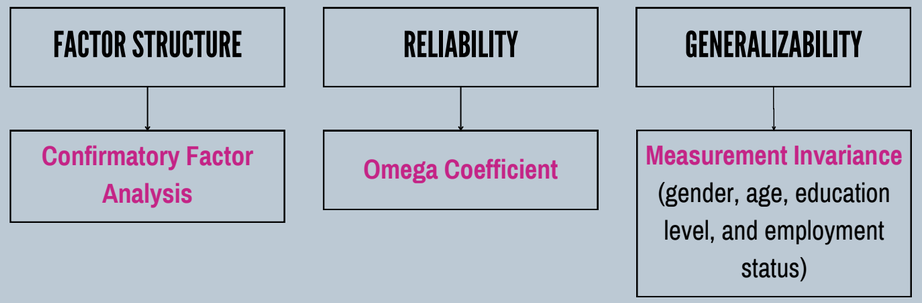
DIFFERENT KINDS OF VALIDITY EVIDENCES through SEM. All models were performed using Mplus. ML, MLR or WLSMV were used as the estimation method, depending on whether the variables were normally distributed, non-normally distributed, or dichotomous, respectively. FIML was used as the method to handle missing data.



AIM & METHODS - DATA ANALYSIS

DESCRIPTIVE STATISTICS using SPSS, to summarize the characteristics of the participants.

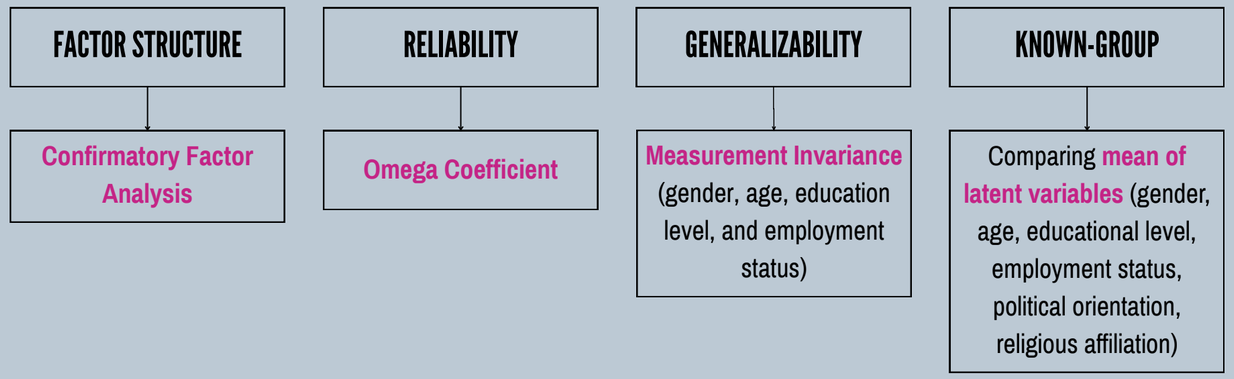
DIFFERENT KINDS OF VALIDITY EVIDENCES through SEM. All models were performed using Mplus. ML, MLR or WLSMV were used as the estimation method, depending on whether the variables were normally distributed, non-normally distributed, or dichotomous, respectively. FIML was used as the method to handle missing data.



AIM & METHODS - DATA ANALYSIS

DESCRIPTIVE STATISTICS using SPSS, to summarize the characteristics of the participants.

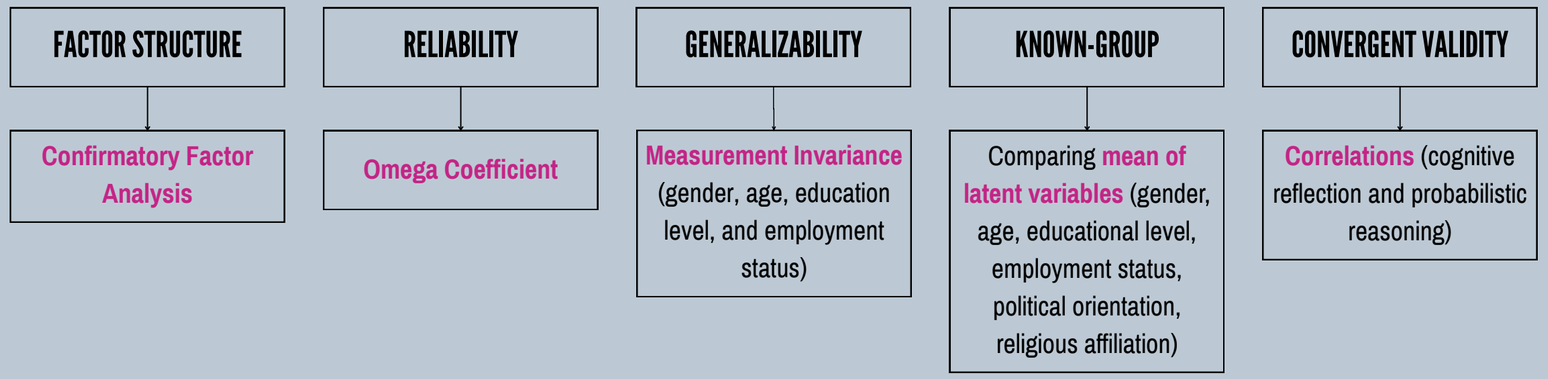
DIFFERENT KINDS OF VALIDITY EVIDENCES through SEM. All models were performed using Mplus. ML, MLR or WLSMV were used as the estimation method, depending on whether the variables were normally distributed, non-normally distributed, or dichotomous, respectively. FIML was used as the method to handle missing data.



AIM & METHODS - DATA ANALYSIS

DESCRIPTIVE STATISTICS using SPSS, to summarize the characteristics of the participants.

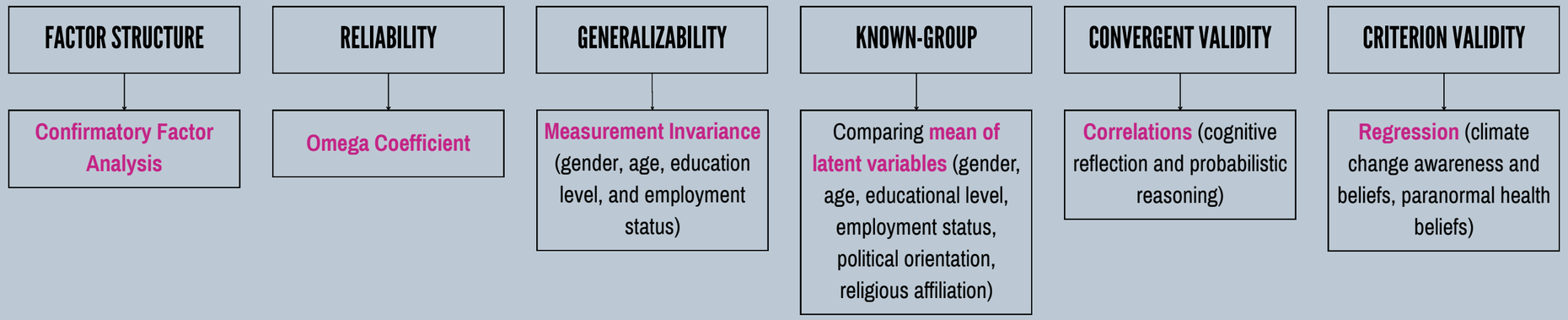
DIFFERENT KINDS OF VALIDITY EVIDENCES through SEM. All models were performed using Mplus. ML, MLR or WLSMV were used as the estimation method, depending on whether the variables were normally distributed, non-normally distributed, or dichotomous, respectively. FIML was used as the method to handle missing data.



AIM & METHODS - DATA ANALYSIS

DESCRIPTIVE STATISTICS using SPSS, to summarize the characteristics of the participants.

DIFFERENT KINDS OF VALIDITY EVIDENCES through SEM. All models were performed using Mplus. ML, MLR or WLSMV were used as the estimation method, depending on whether the variables were normally distributed, non-normally distributed, or dichotomous, respectively. FIML was used as the method to handle missing data.



RESULTS & DISCUSSION

DESCRIPTIVE STATISTICS using SPSS, to summarize the characteristics of the participants.

	N	min	max	M	S.D.
Scientific Reasoning (SRS)	896	0	11	7,71	2,22
Cognitive Reflection (CRT-Long)	891	0	6	2,50	1,94
Probabilistic Reasoning (PRS)	892	1	16	10,25	3,80
Parapsychological beliefs (PHBS, PSIB)	896	1	5	2,15	0,92
Superstitious beliefs (PHBS, SB)	895	1	5	1,45	0,73
Religious beliefs (PHBS, RB)	896	1	5	1,73	0,90
Extraordinary events beliefs (PHBS, EEB)	895	1	5	1,66	0,73
Pseudo-scientific beliefs (PHBS, MEDB)	895	1	5	1,76	0,82
Climate change belifs (CCB)	895	1	7	5,44	1,22
Climate change awareness (CCA)	896	1	7	5,58	1,28
Political orientation (1 item)	897	1	7	3,82	1,62

SRS - % correct responses	
1. Double blind	69,40%
2. Causality	58,10%
3. Confounding variables	74,80%
4. Construct validity	83,00%
5. Control Group	69,50%
6. Ecological validity	72,50%
7. History	67,50%
8. Maturation	75,10%
9. Random assignment to condition	82,10%
10. Reliability	68,40%
11. Response bias	69,40%

RESULTS & DISCUSSION

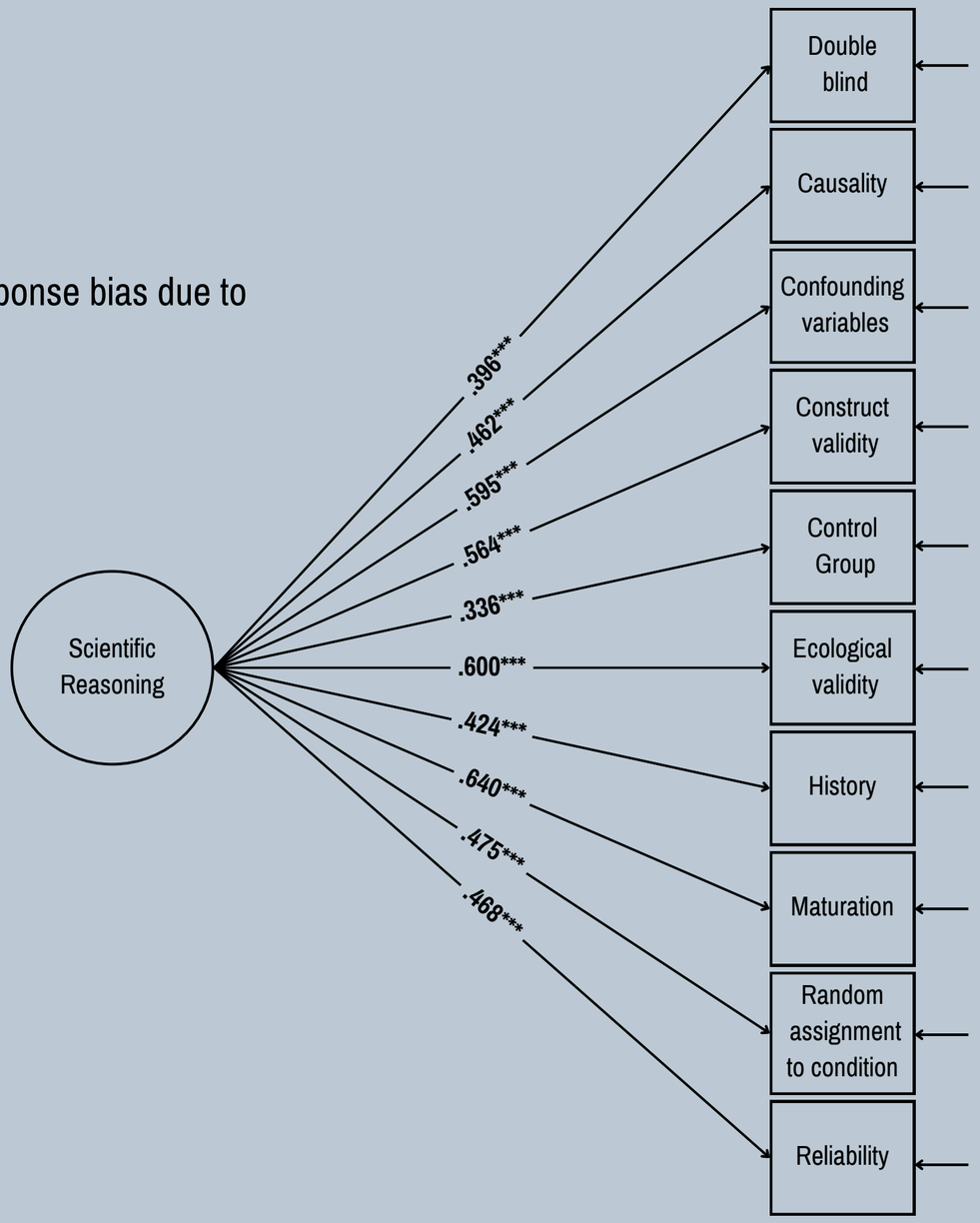
DIFFERENT KINDS OF VALIDITY EVIDENCES through SEM:

1

Via CFA **factor structure evidence** was confirmed, excluding the item on response bias due to insufficient factor saturation.

- The initial fit indices were satisfactory; however, item 11 did not sufficiently saturate the latent factor (loading < .3; Merenda, 1997; Peterson, 2000). Consequently, this item was removed, and a new CFA was conducted using 10 items instead of 11.

- The fit indices for the revised model were good: **[χ^2 (35) = 61.690, p = .004; RMSEA = .029 (.017 .041), p = .999; CFI = .966; WRMR = .953]**. All factor loadings were high (greater than .3) and significant (p < .001), indicating a robust factor structure.



RESULTS & DISCUSSION

DIFFERENT KINDS OF VALIDITY EVIDENCES through SEM:

2

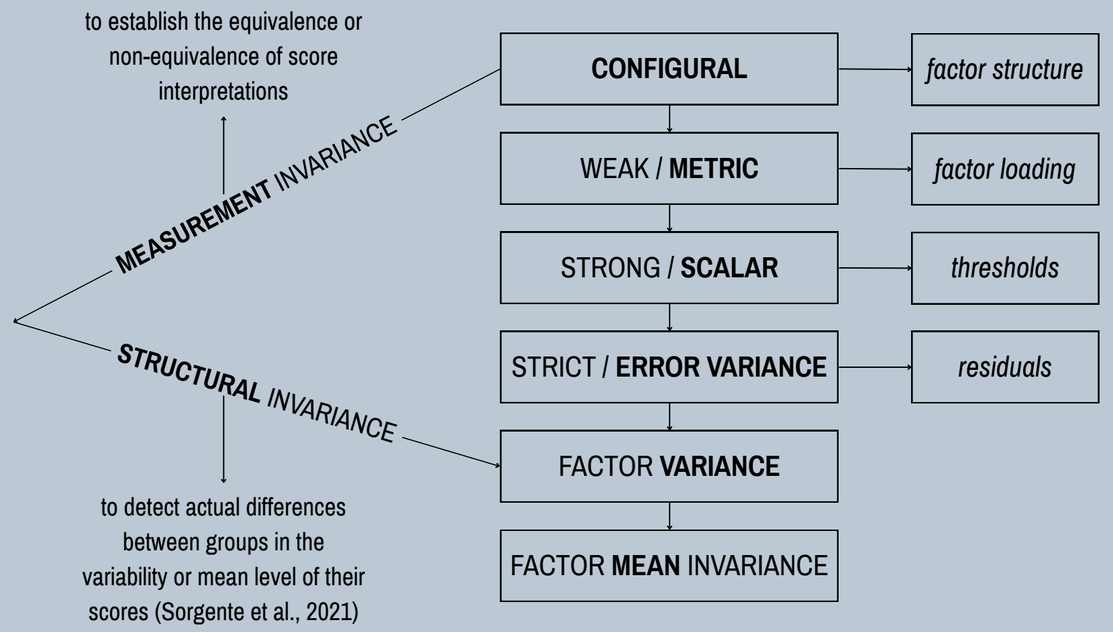
Reliability evidence (ω = .612) confirmed the scale's reliability in the Italian sample.

This value reflects the broad nature of the scientific reasoning construct (Little et al., 1999), which includes 10 distinct content domains.

RESULTS & DISCUSSION

DIFFERENT KINDS OF VALIDITY EVIDENCES through SEM:

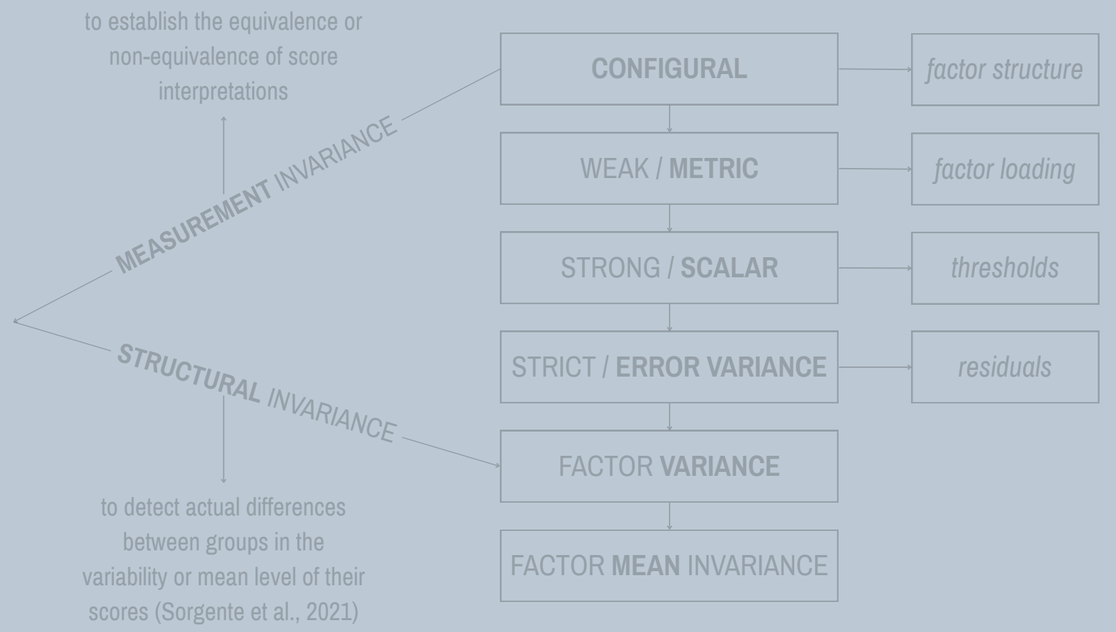
3 **GENERALIZABILITY EVIDENCE** → **Measurement Invariance** → Six types of invariance test were performed:



RESULTS & DISCUSSION

DIFFERENT KINDS OF VALIDITY EVIDENCES through SEM:

3 **GENERALIZABILITY EVIDENCE** → **Measurement Invariance** → Six types of invariance test were performed:



The scale had **fully invariant** **structure, factor loadings, thresholds, residuals, factor variance, and factor means** between:

- males and females (! to achieve strong invariance, it was necessary to free item 9);
- participants aged 18-27 (Gen Z), 28-44 (Gen Y), and 45-60 (Gen X);
- different employment statuses (students, workers, and 'others');
- those who feel politically oriented to the right, center, or left;
- those who feel they belong to a religion and those who do not.

In the comparison based on education level (up to diploma and up to PhD), there was no structural invariance but only measurement invariance, as the difference between the strict and factor invariance models was statistically significant (Difftest $\chi^2 < .001$; $\Delta CFI = -.033$).

→ This full invariance enhances the **scale's applicability** to Italian adults regardless of their background, ensuring that **any observed differences are real** and not due to measurement bias.

RESULTS & DISCUSSION

DIFFERENT KINDS OF VALIDITY EVIDENCES through SEM:

4 **Known-group evidences** was examined by comparing the factor mean of SRS across demographic and ideological groups to determine whether the scale distinguishes between populations anticipated to differ in the measured constructs:

No differences were observed in the factor means of the SRS between:

- MALES AND FEMALE
- AGE GROUPS
- DIFFERENT EMPLOYMENT STATUSES
- POLITICAL ORIENTATION
- RELIGIOUS AFFILIATION

Due to not achieving structural invariance, **differences** in the factor means of the SRS **emerged**:

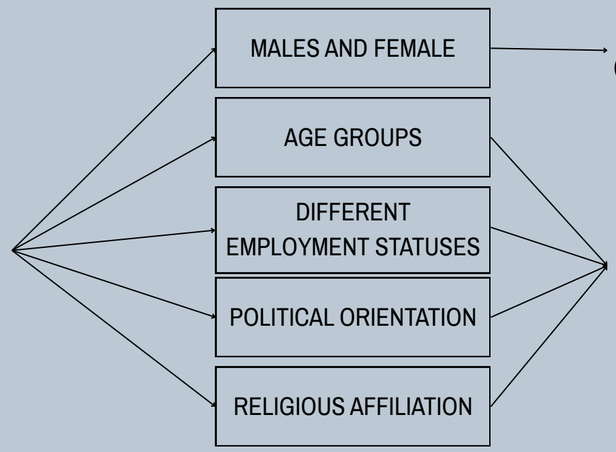
EDUCATIONAL LEVEL → individuals with at least a bachelor's degree had higher factor means than those with at most a high school diploma.

RESULTS & DISCUSSION

DIFFERENT KINDS OF VALIDITY EVIDENCES through SEM:

4 **Known-group evidences** was examined by comparing the factor mean of SRS across demographic and ideological groups to determine whether the scale distinguishes between populations anticipated to differ in the measured constructs:

No differences were observed in the factor means of the SRS between:



Gender-related findings aligned with studies in the U.S., Israel, and Turkey (Drummond & Fischhoff, 2017; Golumbic et al., 2023; Muslu Kaygisiz et al., 2018).

This findings sometimes, but not always, aligned with U.S., Israeli, and Turkish studies. This is unsurprising due to methodological (factor means vs. mean comparisons) and cultural differences (Moyser, 2002).

Due to not achieving structural invariance, **differences** in the factor means of the SRS **emerged**:

EDUCATIONAL LEVEL → individuals with at least a bachelor's degree had higher factor means than those with at most a high school diploma. → suggesting that scientific reasoning skills, as measured by the scale, might be underdeveloped during compulsory schooling and enhanced through university-level training in scientific methodology, regardless of the faculty attended (not only the in STEM fields).

RESULTS & DISCUSSION

DIFFERENT KINDS OF VALIDITY EVIDENCES through SEM:

5

Convergent validity evidence was assessed by correlating the SRS with measures of Cognitive Reflection (CRT-Long) and Probabilistic Reasoning (PRS).

The fit of this model was good:
[χ^2 (53) = 88.14, p = .002; RMSEA = .027 (.017 .037), p = 1; CFI = .973; WRMR = .909]

	Cognitive Reflection Test	Probabilistic Reasoning Scale
SRS	.523***	.578***

This validity is supported by **significant correlations** between the SRS and the CRT and PRS (consistent with Drummond & Fischhoff, 2017).

Thus, higher scientific reasoning scores are associated with higher scores in both cognitive reflection (.523) and probabilistic reasoning (.578), and vice versa.

RESULTS & DISCUSSION

DIFFERENT KINDS OF VALIDITY EVIDENCES through SEM:

6

Criterion-related evidence was examined by testing a SEM model in which the score of SRS was related to different criterion: climate change awareness, climate change beliefs and the five factors of the Paranormal Health Beliefs Scale (parapsychological, superstitious, religious, extraordinary events, and pseudo-scientific).

The fit of the model was good
[χ^2 (98) = 123.85, p = .040); RMSEA = .017 (.004 .026), p = 1; CFI = .989; WRMR = .750]

	SRS
Climate Change Awareness	.118**
Climate Change Beliefs	.131**
Paranormal Health Beliefs	
Parapsychological	-.316***
Superstitious	-.421***
Religious	-.398***
Extraordinary Events	-.406***
Pseudo-scientific	-.394***

RESULTS & DISCUSSION

DIFFERENT KINDS OF VALIDITY EVIDENCES through SEM:

6 **Criterion-related evidence** was examined by testing a SEM model in which the score of SRS was related to different criterion: climate change awareness, climate change beliefs and the five factors of the Paranormal Health Beliefs Scale (parapsychological, superstitious, religious, extraordinary events, and pseudo-scientific).

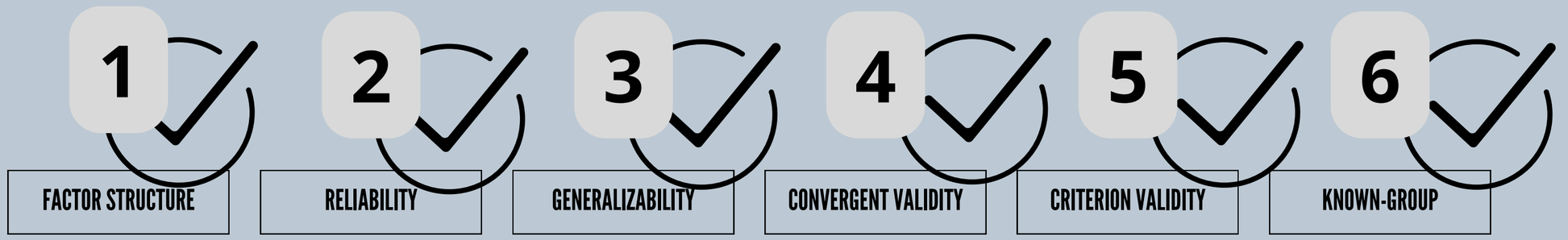
The fit of the model was good
[χ^2 (98) = 123.85, p = .040); RMSEA = .017 (.004 .026), p = 1; CFI = .989; WRMR = .750]

Results showed **significant negative correlations** with five paranormal health belief factors, indicating that higher scientific reasoning scores are associated with fewer paranormal health beliefs. Furthermore, **positive correlations** with climate change awareness and beliefs confirmed that higher scientific reasoning scores are associated with greater climate change awareness and with a stronger tendency to attribute climate change to human activity rather than to natural causes (consistent with Drummond & Fischhoff, 2017).

	SRS
Climate Change Awareness	.118**
Climate Change Beliefs	.131**
Paranormal Health Beliefs	
Parapsychological	-.316***
Superstitious	-.421***
Religious	-.398***
Extraordinary Events	-.406***
Pseudo-scientific	-.394***

RESULTS & DISCUSSION

DIFFERENT KINDS OF VALIDITY EVIDENCES through SEM:



CONCLUSION - KEY FINDINGS

- Successfully validated a revised **Italian version of the SRS** filling a gap in the literature:
  - no previous studies have utilized **SEM** to validate this scale, nor has **measurement invariance** been examined; thus, no assumptions of measurement invariance were made in any country.
  - there is no validated SRS for the Italian context; validating a scale across different cultural contexts is crucial for extending the **construct's cross-national applicability** for providing additional evidence of its validity, which is in keeping with contemporary views of validity.
- Employed **rigorous methodology** (CTT & SEM) ensured strong psychometric properties and provided solid validity evidence consistent with contemporary view of validity;
- Demonstration of **full measurement invariance** across gender, age, education, employment status, political orientation, and religious affiliation strengthens the scale’s generalizability and applicability to the Italian adult population,
- **Education-related differences** suggest a need for formal training in scientific reasoning during compulsory schooling;
- **Strong correlations** with related measures (CRT, PRS) and **predictive power** for paranormal health beliefs and climate change awareness confirm construct validity and practical relevance.

OPEN QUESTIONS & FUTURE RESEARCH

ABOUT THE METHOD...

- **VALIDITY IS NOT SOMETHING THAT IS PROVEN ONCE AND FOREVER IN JUST ONE STUDY**

In line with contemporary view of validity, **future research** could provide further evidence for the scale's validity:

- **ecological validity**: one idea is to create and test a version of the scale for daily life. This could be compared to the current version, which is more focused on lab situations.
- **measurement models**: other studies could also integrate alternative measurement models, such as Item Response Theory; this would allow for triangulated validity evidence regarding the models used to validate the scale;
- **cross-cultural research**: since SRS currently exists in four countries (Italy, US, Turkey, Israel), future research could investigate its cross-cultural invariance, examining whether the construct maintains the same meaning across these diverse cultural contexts.

OPEN QUESTIONS & FUTURE RESEARCH

ABOUT THE METHOD...

- **VALIDITY IS NOT SOMETHING THAT IS PROVEN ONCE AND FOREVER IN JUST ONE STUDY**

In line with contemporary view of validity, **future research** could provide further evidence for the scale's validity:

- **ecological validity**: one idea is to create and test a version of the scale for daily life. This could be compared to the current version, which is more focused on lab situations.
- **measurement models**: other studies could also integrate alternative measurement models, such as Item Response Theory; this would allow for triangulated validity evidence regarding the models used to validate the scale;
- **cross-cultural research**: since SRS currently exists in four countries (Italy, US, Turkey, Israel), future research could investigate its cross-cultural invariance, examining whether the construct maintains the same meaning across these diverse cultural contexts.

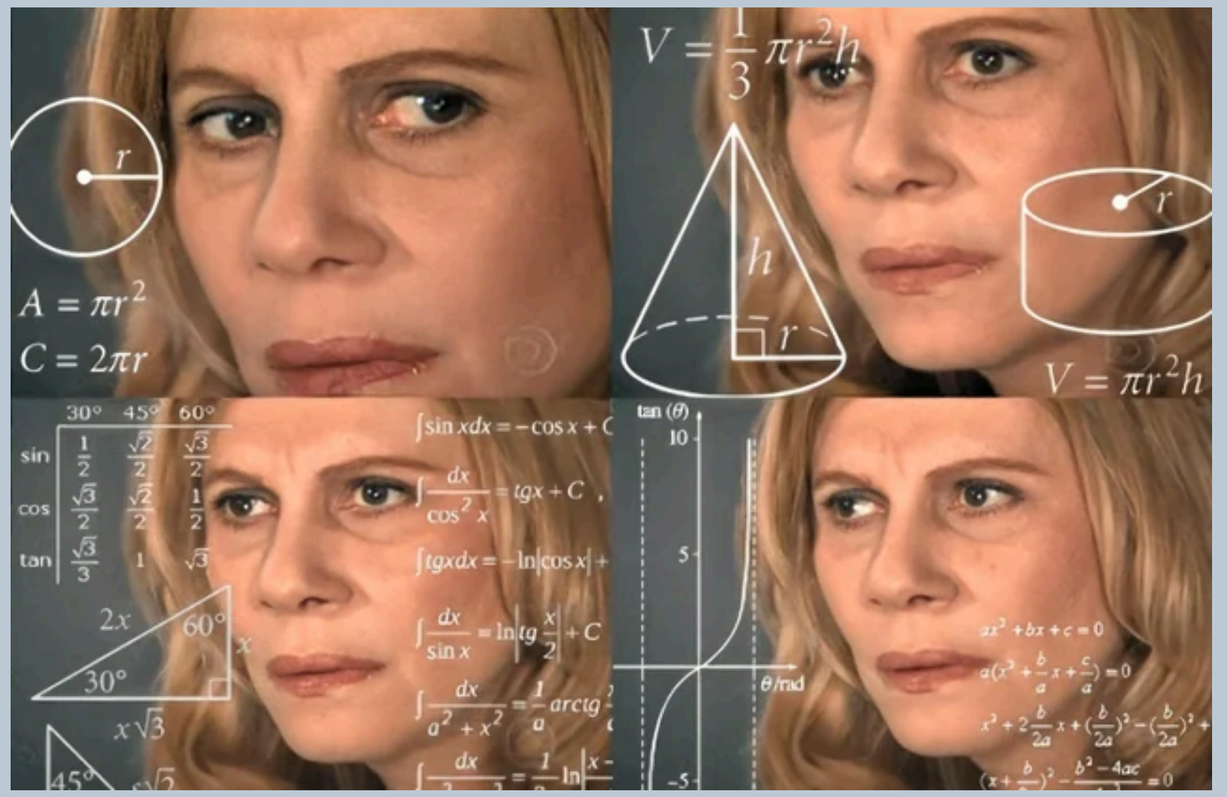
ABOUT THE TOPIC...

- Can the SRS be considered a measure of **scientific reasoning**?
- Does the SRS measure people's ability to think-like-a-scientist in **everyday life**, or is it more like a lab tool designed for scientists?
- Are the 11 concepts measured by the scale really **all the ones we need**?
- If these are the main concepts of scientific reasoning, do we need **three items** for each one?
- There seems to be a difference in **item difficulty**, does that make sense?

FROM EASIEST TO MOST DIFFICULT (BASED ON % OF CORRECT RESPONSES):

- Construct validity
- Random assignment to condition
- Maturation
- Confounding variables
- Ecological validity
- Control group
- Double blind
- Response bias
- Reliability
- History
- Causality

Questions, comments, suggestions, existential doubts – all welcome!



rossella.caliciuri@unicatt.it

DANKESCHÖN

rossella.caliciuri@unicatt.it