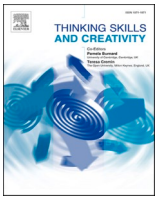
Thinking Skills and Creativity 60 (2026) 102064

Contents lists available at ScienceDirect

Thinking Skills and Creativity

journal homepage: www.elsevier.com/locate/tsc



Scientific reasoning scale in Italy: Validation studies

Rossella Caliciuri *, Margherita Lanz

Università Cattolica del Sacro Cuore, Department of Psychology, Largo A. Gemelli 1, 20123 Milan, Italy

ARTICLE INFO

Keywords:
Scientific reasoning
Classical test theory
Confirmatory factor analysis
Unified view of validity
Contemporary view of validity
Measurement invariance
Structural equation modeling

ABSTRACT

This study presents validity evidence for the Italian adaptation of the Scientific Reasoning Scale (SRS), addressing the lack of a measure of scientific reasoning in the Italian context. A multi-study, multi-method approach was employed, including back-translation, pilot testing, expert interviews, cognitive interviews, and Structural Equation Modeling, to evaluate the psychometric properties and accumulate validity evidence supporting the intended interpretation and use of the scale scores. A total of 897 Italian adults (aged 18–60) participated in the study. Consistent with the contemporary view of validity, we gathered diverse evidence supporting the scale's validity. Confirmatory Factor Analysis confirmed a unidimensional structure and modest composite reliability was observed, suggesting that future item development could strengthen measurement precision. Multi-group analyses supported full measurement invariance across gender, age, employment status, political orientation, and religious affiliation, enhancing generalizability and reducing measurement bias. Education level was the only variable associated with differences in SRS scores, with higher-educated individuals reporting significantly higher levels. Convergent validity was established through correlations with the Cognitive Reflection Test and the Probabilistic Reasoning Scale. Criterion-related validity was demonstrated through relationships with climate change awareness and beliefs, and different paranormal health beliefs. Given the adaptation of the response format from true/false to multiple-choice, this Italian version should be considered an adaptation that, although potentially limiting cross-national comparability, improves linguistic and ecological fit within the Italian context. The Italian SRS provides a valuable tool for future research on scientific reasoning and contributes to the international effort to assess and improve scientific literacy, aligning with the objectives of PISA 2025 ('Programme for International Student Assessment'). The findings regarding the relationship between education level and scientific reasoning scores highlight a potential area for educational intervention, suggesting that formal training in scientific methodology may be necessary to fully develop these skills during compulsory schooling.

1. Introduction

1.1. Scientific reasoning

Scientific reasoning is defined as the investigative skills used in scientific inquiry to build and refine knowledge. This encompasses the systematic exploration of a problem, the formulation and testing of hypotheses, the manipulation and isolation of variables, and the

* Corresponding author at: Università Cattolica del Sacro Cuore, Largo A. Gemelli 1, 20123 Milan, MI Italy.
E-mail addresses: rossella.caliciuri@unicatt.it (R. Caliciuri), margherita.lanz@unicatt.it (M. Lanz).

https://doi.org/10.1016/j.tsc.2025.102064
Received 1 April 2025; Received in revised form 28 October 2025; Accepted 13 November 2025
Available online 13 November 2025

observation and evaluation of consequences (Bao et al., 2009, 2022; Johnson & Lawson, 1998). Despite its significance, a paucity of consensus persists with regard to its definition and operationalisation. Diaz et al. (2023) highlight these inconsistencies in their conceptual review, defining scientific reasoning as a set of specialised cognitive processes within the realm of thinking. These processes include induction (Eysenck & Keane, 2020), deduction (Garnham & Oakhill, 1994), and abduction (Johnson-Laird & Byrne, 1993). While the terms 'scientific method' and 'thinking like a scientist' are often used to define scientific reasoning (Diaz et al., 2023), they do not fully encompass its cognitive underpinnings. The scientific method refers to a structured series of steps to understand phenomena, whereas 'thinking like a scientist' involves applying these steps to inquiry (Diaz et al., 2023; Kisiel et al., 2012). Scientific reasoning, therefore, comprises the cognitive processes necessary to implement the scientific method. A widely accepted model by Fischer et al. (2014) describes scientific reasoning as a sequential process including (a) problem identification, (b) question formulation, (c) hypothesis generation, (d) artifact construction, (e) evidence generation, (f) evidence evaluation, (g) conclusion drawing, and (h) result communication. There are, therefore, different definitions of scientific reasoning, as well as divergent conceptualisations of its development. These reflect a divergence in opinions on the relative merits of conceptual knowledge versus strategic acquisition in explaining developmental differences. As Zimmerman (2000) demonstrates, the emphasis placed on conceptual knowledge varies between instruments, with some placing greater emphasis on this aspect, whilst others prioritise experimental strategies. In general, most existing measures are tailored to specific populations and contexts, limiting their generalizability.

One notable attempt to measure scientific reasoning is the Scientific Reasoning Scale (SRS; Drummond & Fischhoff, 2017), which was originally developed in the US and subsequently adapted for use in Turkey (Muslu Kaygisiz et al., 2018). The SRS assesses individuals' ability to evaluate scientific evidence through an interdisciplinary approach incorporating behavioral decision research, cognitive developmental psychology, and public understanding of science. The scale consists of 11 true/false items that challenge participants to apply reasoning skills to brief scientific scenarios, reflecting a variety of facets of scientific reasoning, such as blind/double-blind experiments, causality, confounding variables, construct validity, control group, ecological validity, history, maturation, random assignment to conditions, reliability, and response bias. The scale is intended to function as a unidimensional measure of scientific reasoning, and therefore the total score reflects the number of correct responses. This unidimensional scale can be considered 'lab-oriented', as the scenarios presented represent typical laboratory research scenes. Despite its usefulness, no validity evidence for the SRS has yet been gathered in Italy, where no measure of scientific reasoning currently exists. Having measures supported by validity evidence is essential for several reasons. First, it facilitates meaningful comparisons of scientific reasoning across different countries, reflecting how well individuals are prepared to engage with scientific concepts and issues. This objective is consistent with the overarching aims of the PISA 2025 assessment, which is administered by the OECD (Organisation for Economic Co-operation and Development). Second, from a psychometric standpoint, and in line with contemporary views of validity (Hubley & Zumbo, 2011), providing validity evidence for a measurement tool across different cultural contexts provides additional evidence for its construct validity, particularly concerning the generalizability of its structure and function across populations. An essential part of this process is testing measurement invariance, which allows researchers to evaluate whether the instrument operates equivalently across groups, ensuring that observed differences in scores reflect true differences in the construct rather than artifacts of the measurement process. Furthermore, construct validation, as originally outlined by Cronbach and Meehl (1955), and more recently revisited by Flake and Fried (2020), is a fundamental step in psychological science, as it ensures that instruments genuinely measure the constructs they are intended to assess.

1.2. *Contemporary view of validity*

In the latest edition of the 'Standards for Educational and Psychological Testing' (American Educational Research Association [AERA], American Psychological Association [APA], & National Council on Measurement in Education & [NCME], 2014), the contemporary view of validity, also referred to as the 'unified theory of validity', is positioned as the cornerstone of test validation. This modern conceptualization, primarily advanced by Messick (1989, 1995), defines validity not as a static property of a test but as a holistic and ongoing process of gathering multiple sources of evidence to support the interpretation and use of test scores. It integrates previously distinct types of validity (e.g., content, criterion, construct) into a single overarching framework, construct validity, and emphasizes that all validity is construct validity. In this framework, validity is considered to be context- and sample-dependent, with test developers and users jointly responsible for providing evidence that justifies score interpretation. According to this unified model, various sources of evidence contribute to the validation process, including content validity, criterion-related validity (both concurrent and predictive), factorial structure, convergent and discriminant validity, generalizability, known-group evidence, and consequences of testing. The latter concerns how the anticipated and unanticipated social or individual outcomes of test use confirm its appropriateness and quality; this form of evidence is often gathered in follow-up studies (Cronbach & Meehl, 1955; Hubley & Zumbo, 2011; Sorgente & Zumbo, 2025; Zumbo, 2005). In this validation process, the use of SEM is pivotal, as they integrate regression, path analysis, and latent variable models (factor analysis).

1.3. *The present study*

The present study aims to provide validity evidence for the Italian adaptation of the SRS. A multi-study, multi-method approach was employed, as suggested by DeVellis and Thorpe (2021) and in line with other studies (de Oliveira Cardoso et al., 2024; Sorgente & Lanz, 2019). Study 1 involved backtranslating the scale, pilot testing, and preliminary confirmatory factor analysis (CFA). Study 2 consisted of qualitative and quantitative phases. The qualitative phase used expert interviews to evaluate item adequacy and cognitive interviews to test comprehensibility. The quantitative phase 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 the contemporary view of validity (Hubley & Zumbo, 2011). Specifically, the following sources of evidence will be examined: (1) Factorial structure evidence through CFA, verifying whether the observed variables align with the hypothesized latent factors and confirming the underlying structure of the scale. (2) Reliability evidence by calculating the omega coefficient to assess the internal consistency and precision of the scale in measuring the intended constructs. (3) Generalizability and known-group evidences by testing measurement invariance and comparing latent mean differences across demographic and ideological groups (i.e., gender, age, education level, employment status, political orientation, and religious affiliation) to assess whether the scale functions equivalently and distinguishes between populations expected to differ on the measured constructs. (4) Convergent validity evidence by analyzing correlations with theoretically related constructs (cognitive reflection and probabilistic reasoning) to determine the extent to which the scale aligns with related measures. (5) Criterion validity evidence by using SEM-based regression to examine the relationship between the scale and external criteria (climate change awareness; climate change belief; and five factors of Paranormal Health Beliefs Scale: parapsychological beliefs, superstitious beliefs, religious beliefs, extraordinary events beliefs, and pseudo-scientific beliefs). This step assesses how well the scale predicts these relevant outcomes. Fig. 1 presents the structure of the multi-study and multi-method project conducted to provide validity evidence for the SRS. Given their preliminary nature, Study 1 and the qualitative phase of Study 2 are described briefly. In contrast, the quantitative phase is reported in detail in the following sections.

1.3.1. Study 1

In the first study, we back-translated and adapted the SRS items into Italian and then conducted pilot testing to provide initial validity evidence for the Italian adaptation of the scale, as suggested by the guidelines of Kowal (2024) and DeVellis and Thorpe (2021). Following administration of the SRS to 337 Italian adults, we assessed the scale's factor structure and identified several psychometric weaknesses. Specifically, the CFA revealed several issues in the scale's structure. Five items did not reach the minimum recommended factor loading threshold (e.g., loading $\geq$ 0.30; Merenda, 1997; Peterson, 2000), indicating weak alignment with the theoretical construct. Items 1, 3, 5, 7 e 11 showed loadings below this cut-off. These findings raised concerns about both the linguistic adequacy of the Italian adaptation and the clarity and effectiveness of the original true/false response format. These limitations provided the rationale for conducting Study 2. The details of the study 1 (preregistration, dataset, codebook, codes, and results) are openly available on OSF [Open Science Framework] and referenced in the 'Data publicly available in a repository'.

1.3.2. Study 2

Following the results of the first study, and as part of the qualitative phase, both the linguistic formulation of the items and the response format of the scale were revised. The original SRS employed a dichotomous true/false format embedded within the item stem. In the present adaptation of the scale, two discrete multiple-choice response options were implemented for all 11 items. To illustrate, consider the item addressing double-blind facet. The original item presented the scenario with a true/false question embedded within the description: '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'. The revised item presents the same scenario, but the response options are distinct: 'In an experiment to evaluate coffee taste preferences, a researcher places coffee from brand A and coffee from brand B in identical cups. To differentiate the types of coffee, the researcher puts identifying labels at the bottom of the cups. Subsequently, a laboratory assistant distributes one of the cups to the tasters while the researcher observes their facial expressions'. The response options are: 'a. The laboratory assistant should not look at the cups while they are being filled' and 'b. The laboratory assistant should look at the cups while they are being filled'. This decision was based on several considerations: (1) feedback from cognitive interviews indicating confusion due to the embedded true/false phrasing; (2) the need to increase clarity and reduce ambiguity in participants' processing of the response task; and (3) alignment with evidence from response process theory, which suggests that separating the scenario from the response alternatives helps elicit more deliberate and interpretable cognitive operations (Zumbo & Hubley, 2017). This modification represents a substantive adaptation rather than a mere translation of the original scale. While preserving the conceptual content of the items, the new response format improves ecological and linguistic fit for Italian respondents and reduces response ambiguity. However, it also limits the degree of direct cross-national comparability with studies using the original

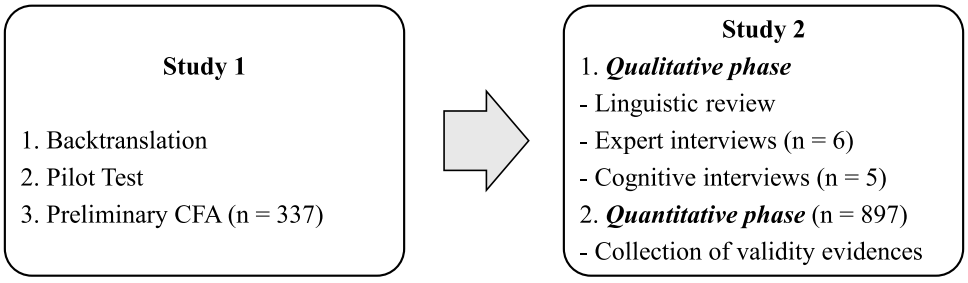


Fig. 1. Structure of the multi-study and multi-method project conducted to provide validity evidence for the scientific reasoning scale (SRS).

dichotomous format.

In order to revise the items and gain further insight into the construct within the specific Italian context, six expert interviews were conducted. The expert interviews were semi-structured and followed a protocol consisting of three main sections: (1) evaluation of item clarity and linguistic precision; (2) assessment of the alignment between each item and the intended construct of scientific reasoning; and (3) suggestions for improved wording, response format, or both cultural and construct-related appropriateness. The composition of the 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). Each expert was presented with the items, and invited to comment on ambiguities or cultural incongruities. Based on their feedback, all items were reviewed and revised. For instance, one expert advised against the use of overly definitive adverbs such as 'necessarily' in response options, as these may imply absolute conclusions that are not in accordance with scientific reasoning. This resulted in a first draft of the scale, whose comprehensibility and applicability to the lives of Italian adults was tested through five cognitive interviews with the target population (the participants were selected to represent a range of genders, age ranges and educational backgrounds). Cognitive interviews are a specific type of structured interview focused on the cognitive processes respondents utilize when answering 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). After answering each item, participants were asked to reflect aloud on how they interpreted the question, what they believed it was asking, and how they arrived at their answer. When misinterpretations occurred, we made targeted revisions to improve clarity. For instance, a participant reported difficulty understanding the purpose of item 1 due to the abstract wording, prompting a revision using a more concrete example. After evaluating the adequacy and the comprehensibility of the items, we proceeded with the quantitative phase to assess the psychometric properties of the scale, providing validity evidence in accordance with the contemporary view of validity.

2. Method

2.1. Participants

The data was collected from December 2024 to January 2025. Participants were recruited through quota sampling, a non-probability technique designed to approximate population representativeness by stratifying the population by key sociodemographic variables (age, gender, and geographical region) and setting quotas proportional to national distributions. The sample consists of 897 Italian participants: 50.50 % identify as female, 48.94 % as male and the remaining 0.56 % identify as non-binary or prefer not to disclose their gender. The age of the participants ranges from 18 to 60 years, with a mean age of 41.48 years (SD = 12.46; median = 43). Nearly all participants were born, resided and completed their education entirely in Italy (96.32 %, 99.44 % and 97.76 %, respectively). Further information regarding the sample characteristics is presented in Table 1. Although not based on probability sampling, the quota design ensured that the distribution of key demographic variables closely reflected the Italian population aged 18–60 years, according to ISTAT estimates for 2024 (Istituto Nazionale di Statistica, 2024). This choice was made to ensure that the validity evidence would be applicable to the broader population for which the instrument is intended. As highlighted in the validation literature, "validity is not a property of the test itself but of the inferences made from the test scores, and these inferences must be justified within the population and context of use" (Zumbo & Hubley, 2017, p. 19). Adopting a demographically balanced sample enhances the generalizability and ecological validity of the results and aligns with best practices in psychometric validation.

Table 1
Demographic characteristics of Italian participants ($N = 897$).

Characteristic	%
Age Groups	
Ages 18 – 27 (genZ)	20.18
Ages 28 – 44 (genY)	35.12
Ages 45 – 60 (genX)	44.70
Region of Residence	
Northwestern Italy	25.65
Northeastern Italy	19.57
Central Italy	20.58
Southern Italy	22.95
Italian Islands	11.25
Education level	
Middle School	7.28
High School Diploma	52.86
Bachelor's/Master's Degree or equivalent	36.28
Doctoral Degree	3.58
Employment Status	
Students	15.46
Employed	67.20
Unemployed	17.37

2.2. Procedure

The project was preregistered on OSF in December 2024. The protocol was approved by the Ethical Committee of the Department of Psychology of Università Cattolica del Sacro Cuore of Milan and adhered to all principles of the Declaration of Helsinki. The study utilized an online survey administered from December 2024 to January 2025. The sample consisted of 897 respondents, randomly selected from the consumer panel managed by Norstat (https://norstat.it/). In order to ensure that the sample reflected the adult Italian population, quotas were established in accordance with ISTAT statistics, which are also available on OSF. To improve data quality, we included some attention-check questions in the questionnaire and implemented *Google Invisible ReCaptcha technology* to identify and potentially exclude inattentive individuals and/or bots. All 897 participants successfully passed these quality control checks. Participation in the survey was entirely voluntary and was compensated according to the platform's guidelines. To access the questionnaire and participate in the study, informed consent was mandatory. In addition to providing informed consent, participants also agreed to share their data on open science platforms, making their responses publicly available for research. The online questionnaire, administered through Qualtrics as a single survey session, took approximately 45 min to complete.

2.3. Instruments

The participants initially provided socio-demographic information, including gender, age, region of residence, education level, and employment status.

To investigate participants' subjective perception of how much they view their activities—whether in their job, school, or field of study – as connected to science, we posed ad hoc question: 'Do you think your job/school/faculty is related to science?'. Participants had three response options: 'Yes, my job/school/faculty is closely related to science', 'My job/school/faculty has some connections with science', and 'No, my job/school/faculty is not directly related to science', with the option to provide a rationale for their response.

To measure scientific reasoning, we administered our Italian adaptation of the SRS (Drummond & Fischhoff, 2017), a dichotomous scale consisting of 11 items. The Italian SRS can be found on OSF and in Appendix A.1, while the details on how we developed this adaptation are discussed in the introduction.

In line with the original development and validation work on the SRS, scales theoretically and empirically linked to scientific reasoning were administered, allowing for both consistency across studies and construct coherence (Drummond & Fischhoff, 2017; Golumbic et al., 2023; Muslu Kaygisiz et al., 2018). In particular, regarding the convergent measures, we administered two measurement scales: The Cognitive Reflection Test-Long (CRT-Long; Primi et al., 2016) is an extended version of the CRT (Frederick, 2005) is an extended version of the original CRT (Frederick, 2005). While the original CRT consisted of 3 questions, the CRT-Long includes 6 questions. Performance on the CRT is associated with cognitive ability and with the capacity to resist intuitive but incorrect responses. In this sense, it serves as a measure of reflective, rational thinking and reduced susceptibility to cognitive biases (Toplak et al., 2011). An example item is the following: 'If three elves can wrap three toys in one hour, how many elves are needed to wrap six toys in two hours?' [correct answer = 3 elves; heuristic answer = 6 elves]. Participants' scores are determined based on the sum of correct answers to these items. The unidimensional structure of the scale was tested with a CFA. Fit indices for this model were good [χ^2 (9) = 22.36, p = .008); RMSEA = 0.04 (0.02 - 0.06); CFI = 0.996; WRMR = 0.80]. The factor loadings were all high (> |.71|) and significant (p < .001). Internal consistency score was good (McDonald's ω = 0.80).

The Probabilistic Reasoning Scale (PRS; Primi et al., 2017) is a measure of basic probabilistic reasoning skills that are necessary to successfully interpret probability information in everyday settings, as well as to complete introductory statistics courses. The scale consists of 16 items. An example item is the following: 'A ball is drawn from an urn containing 10 red, 20 blue, 30 white, and 15 yellow balls. What is the probability that it is neither red nor blue?'. The answer choices are: 'a) 30/75; b) 10/75; c) 45/75′. Participants' scores are determined based on the sum of correct answers to these items. The unidimensional structure of the scale was tested with a CFA. Fit indices for this model were good [χ^2 (104) = 307.10, p < .001); RMSEA = 0.05 (0.04 - 0.05); CFI = 0.954; WRMR = 1.34]. The factor loadings were all high (> |.30|) and significant (p < .001). Internal consistency score was good (McDonald's ω = 0.81).

Regarding the outcome measures, we have the following scales or items: The Paranormal Health Beliefs Scale (PHBS, Donizzetti & Petrillo, 2017) is an instrument for the assessment of the range of illusory beliefs about health, through a series of 31 items. This scale consists of 5 dimensions: religious belief (RB; 8 item, e.g., 'There are saints who may cure and protect the health of the body '), superstitious belief (SB; 7 item, e.g., 'Preferably avoid surgical interventions on Friday 17th '), belief in extraordinary events (EEB; 6 item, e.g., 'The appearance of a new disease may be due to extra-terrestrial entities '), parapsychological belief (PSIB; 6 item, e.g., 'Illness can be overcome by force of mind'), and pseudo-scientific of a biomedical nature belief (MEDB; 4 item, e.g., 'There are some social groups, e.g., immigrants, who adopt unhealthy habits that put the health of the human species in danger'). Participants rated their agreement with statements using a Likert scale ranging from 1 ('Completely Disagree') to 5 ('Completely Agree'). The multi-dimensional structure of the scale was tested with a CFA. Fit indices for this model were good [χ^2 (424) = 1209.04, p < .001); RMSEA = 0.05 (0.04 - 0.05); CFI = 0.925; SRMR = 0.05]. The factor loadings were all high (> |.77|) and significant (p < .001). Internal consistency scores ranged from acceptable to excellent, with McDonald's ω = 0.73 (MEDB), 0.81 (EEB), 0.85 (PSIB), 0.91 (SB), and 0.93 (RB).

Climate Change Beliefs: participants' beliefs regarding climate change will be measured using two items commonly used in research on the issue (Bertolotti et al., 2021): 'As far as you know, do you personally think that the world's climate is changing?', on a 7-point scale ranging from 1 ('Not at all') to 7 ('Very much'), and 'Do you think that climate change is caused by natural processes or by human activity?', on a scale ranging from 1 ('Natural processes only') to 7 ('Human activity only'). These two items are adequately

correlated with each other (Spearman $\rho = 0.476$).

Climate Change Awareness: individuals' concern for climate change issues will be measured through six items used in Robba et al. (2024) developed on a 7-point Likert scale from 1 ('I totally disagree') to 7 ('I totally agree'). An example of item is the following: 'Climate change is pushing the planet to a point of no return'. The unidimensional structure of the scale was tested with a CFA. Fit indices for this model were good [χ^2 (8) = 30.75, $p < .001$); RMSEA = 0.06 (0.04 - 0.08), $p = .278$; CFI = 0.986; SRMR = 0.17]. The factor loadings were all high ($> |.79|$) and significant ($p < .001$). Internal consistency score was excellent (McDonald's $\omega = 0.93$).

In line with the original SRS construction study (Drummond & Fischhoff, 2017), we included an item to assess political orientation: 'Indicate your political orientation on a 7-point scale ranging from left (1) to right (7)', sourced from an existing study (Di Battista et al., 2018), as well as religious affiliation using a dichotomous item: 'Do you feel you belong to a religion?' (response options: 'Yes/No').

Further detailed information regarding the variables and scales incorporated within the survey, in addition to information on informed consent, data sharing, instructions, and the entirety of the items contained within the research's codebook, can be found among the materials uploaded on the OSF.

2.4. Data analysis

Descriptive statistics. Descriptive statistics, including means and standard deviation, were used to summarize the characteristics of the participants using SPSS (version 29.0.2.0; IBM Corp., 2023).

As suggested by Zumbo (2005), we tested different kinds of validity evidence (factor structure evidence, generalizability evidence, convergent evidence, criterion-related evidence, and reliability evidence) through structural equation modeling. All models were performed using Mplus software (version 7.11; Muthén & Muthén, 1998-2017). Maximum Likelihood (ML), Maximum Likelihood Robust (MLR), or Weighted Least Squares Mean and Variance Adjusted (WLSMV) were used as the estimation method, depending on whether the variables were normally distributed, non-normally distributed, or dichotomous, respectively. Full Information Maximum Likelihood (FIML) was used as the method to handle missing data.

Factor structure and reliability evidences. After describing the data, for each scale adopted (SRS, CRT, PRQ, PHBS, CCA), we tested the factorial structure with CFA and saved their factor scores for further analysis. In order to test the goodness of fit of the CFA model, we considered the exact fit index, represented by the χ^2 value, and several approximate fit indices, including the Root Mean Square Error of Approximation (RMSEA), the Comparative Fit Index (CFI), or the Tucker–Lewis Index (TLI), the Standardized Root Mean Square Residual (SRMR), or the Weighted Root Mean Square Residual (WRMR). A non-significant χ^2 value suggests that the model is consistent with the data, although this index is highly sensitive to sample size (Cheung & Rensvold, 2002). RMSEA values close to zero indicate a better fit, with values below 0.08 considered reasonable and those below 0.05 considered good (Lai & Green, 2016; Kline, 2023). CFI and TLI values close to 1 suggest a good model fit, whereas values below 0.90 indicate a poor fit (Lai & Green, 2016). SRMR values close to zero reflect a better fit, with values below 0.08 indicating an adequate fit (Hu & Bentler, 1999). Finally, WRMR values below 1.0 are generally interpreted as evidence of an acceptable model fit (Asparouhov & Muthén, 2018). However, these interpretation guidelines related to goodness-of-fit indexes were not treated as 'golden rules' or used for inferential purposes, but only as rough guidelines for descriptive model evaluation, to integrate with parameter estimates, statistical conformity, and theoretical adequacy (Fan & Sivo, 2007). After confirming the adequacy of the models in terms of factorial structure, the reliability of the scales was assessed. Following current guidelines (Dunn et al., 2014), internal consistency was estimated using composite reliability (ω).

Generalizability evidence. The generalizability evidence for the SRS was tested using measurement invariance. A multi-group analysis was conducted, with groups defined by gender, age, education level, or employment status, political orientation and religious affiliation. These groups were compared according to four types of measurement invariance: configural, weak, strong, and strict invariance. The levels were tested sequentially, from the least restrictive (configural) to the most restrictive (strict). To determine whether a specific level of invariance was met, each model was compared to a less constrained version. Significant differences between models were evaluated using the chi-square difference test, which assesses whether constraints imposed on the model significantly alter its fit. However, in large samples ($n > 300$), chi-square tests can be overly sensitive, detecting even minor and trivial differences. Therefore, model comparisons often rely on practical significance, assessed through approximate fit indices. In this study, we used changes in the Comparative Fit Index (ΔCFI), where a decrease of 0.010 or more (Cheung & Rensvold, 2002; Meade et al., 2008) indicates measurement non-invariance. Additionally, changes in other fit indices (TLI and RMSEA) were examined. If full invariance was not achieved at a particular level, we tested for partial invariance to identify specific parameters that did not hold across groups. Following Saris et al. (2009), parameters were freed based on the standardized expected parameter change reported in the Mplus output. According to Dimitrov (2010), full invariance across all items is not required to consider a scale sufficiently equivalent across groups. If at least 80 % of the items demonstrate full measurement invariance, the instrument can be deemed valid for cross-group comparisons. Specifically, achieving weak measurement invariance (equal factor loadings) allows for the comparison of total factor variability across groups. Achieving strong measurement invariance (equal intercepts/thresholds) further enables comparisons of factor means. When strong measurement invariance is established, any observed differences in factor variability or mean levels can be interpreted as genuine differences in the underlying construct rather than artifacts of measurement bias. In order to verify if the SRS latent factor's variability as well as mean level were significantly different across groups, we constrained respectively the SRS variance and the SRS mean to be equivalent across groups and then verified if this constraint significantly modified the model fit ($\Delta CFI < -.010$). This statistical procedure is called 'structural invariance' (Widaman et al., 2014). All in all, six types of invariance test were performed (configural, weak, strong and strict, factor variance, factor mean invariance) for each group comparison. The first four types correspond to the measurement invariance, while the last two types correspond to the structural invariance. While measurement

invariance is designed to help establish equivalence/nonequivalence of score interpretations, structural invariance is designed to detect actual differences between/among groups in the variability or mean level of their scores (Sorgente et al., 2021). When we compared latent parameters (factors variance and mean), we adopted the chi-square difference test as the change in CFI is not sensitive enough for meaningful change on them. At the same time, as the χ^2 statistic is overly sensitive for large numbers of constraints, especially when estimated on large sample sizes (e.g., Marsh et al., 1988), we did not use the cut-off p=.05 to consider the two compared models significantly different. Specifically, we considered an adjusted p-value of 0.001 in order to reduce the possibility of type I error (Little, 1997). In total, we conducted four invariance tests for each group comparison: configural, weak, strong, and strict. Measurement invariance assesses whether score interpretations remain consistent across groups, ensuring that the instrument functions reliably in different populations.

Known-group evidence. Known-group evidence was examined by comparing mean of SRS across different demographic and ideological groups to evaluate whether the scale effectively distinguishes between populations expected to differ on the measured constructs (based on gender, age, educational level, employment status, political orientation, religious affiliation).

Convergent evidence. Convergent validity was assessed by correlating the SRS factors with measures of cognitive reflection and probabilistic reasoning.

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

The dataset is available on the OSF page, as well as the SPSS syntax and the Mplus input files.

3. Results

Descriptive statistics. Demographic information is presented in the 'Participants' section. On average, participants answered correctly to 7.71 out of 11 SRS items (SD = 2.22; median = 8). Table 2 displays the percentage of correct responses for each item of the scale as provided by the sample. The scores range from 58.10 % (Causality) to 83.00 % (Construct Validity).

Regarding religious affiliation, 52.8 % of the sample ($n = 474$) feels a sense of belonging to a religion, while the remaining 47.2 % ($n = 423$) does not.

Descriptive statistics (means and standard deviations) for all other scales included in the survey are available in the supplementary materials on OSF.

Factor structure and reliability evidences. The factor structure of the Italian SRS was tested using CFA on a sample of 896 participants. The initial fit indices were satisfactory; however, item 11 did not sufficiently saturate the latent factor (loading < 0.3; Merenda, 1997; Peterson, 2000). Consequently, this item was removed, and a new CFA was conducted using 10 items. The fit indices for the revised model were good: [χ^2 (35) = 61.690, $p = .004$; RMSEA = 0.029 (0.017 0.041), $p = .999$; CFI = 0.966; WRMR = 0.953]. As shown in Fig. 2, all factor loadings were high (< 0.3) and significant ($p < .001$), all ten items load significantly on a single latent factor representing scientific reasoning, with standardized loadings ranging from 0.336 to 0.640. These results confirm the unidimensionality of the scale. Composite reliability was estimated for SRS: $\omega = 0.612$.

Generalizability and known-group evidences. Multi-group analyses were performed in order to collect evidence about the generalizability of the interpretation of the test scores across different subgroups. Specifically, seven types of invariances (configural, metric, scalar, uniqueness, factor variance, factor covariance, and factor mean; Bontempo et al., 2007) were tested across variables relevant to scientific reasoning: gender, age, education level, employment status, political orientation, and religious affiliation (see Table 3 for all model comparisons and fit indices). The results showed that the scale demonstrated full invariance in structure, factor loadings, thresholds, residuals, factor variances, and factor means across gender, age groups (18–27, 28–44, and 45–60), employment status (students, workers, and others), political orientation (left, center, right), and religious affiliation (believers vs. non-believers). However, for the comparison based on gender, to achieve strong invariance, it was necessary to free item 9 (related to random assignment to condition), for which the probability of providing a correct response was greater in the female group than in the male group. No significant differences in SRS factor means were found across gender, age, employment status, political orientation, or religious affiliation. In contrast, for educational level (up to high school diploma vs. at least bachelor's degree), structural invariance was not

Table 2
Descriptive statistics of SRS.

	% of correct responses ($N = 897$)
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 %

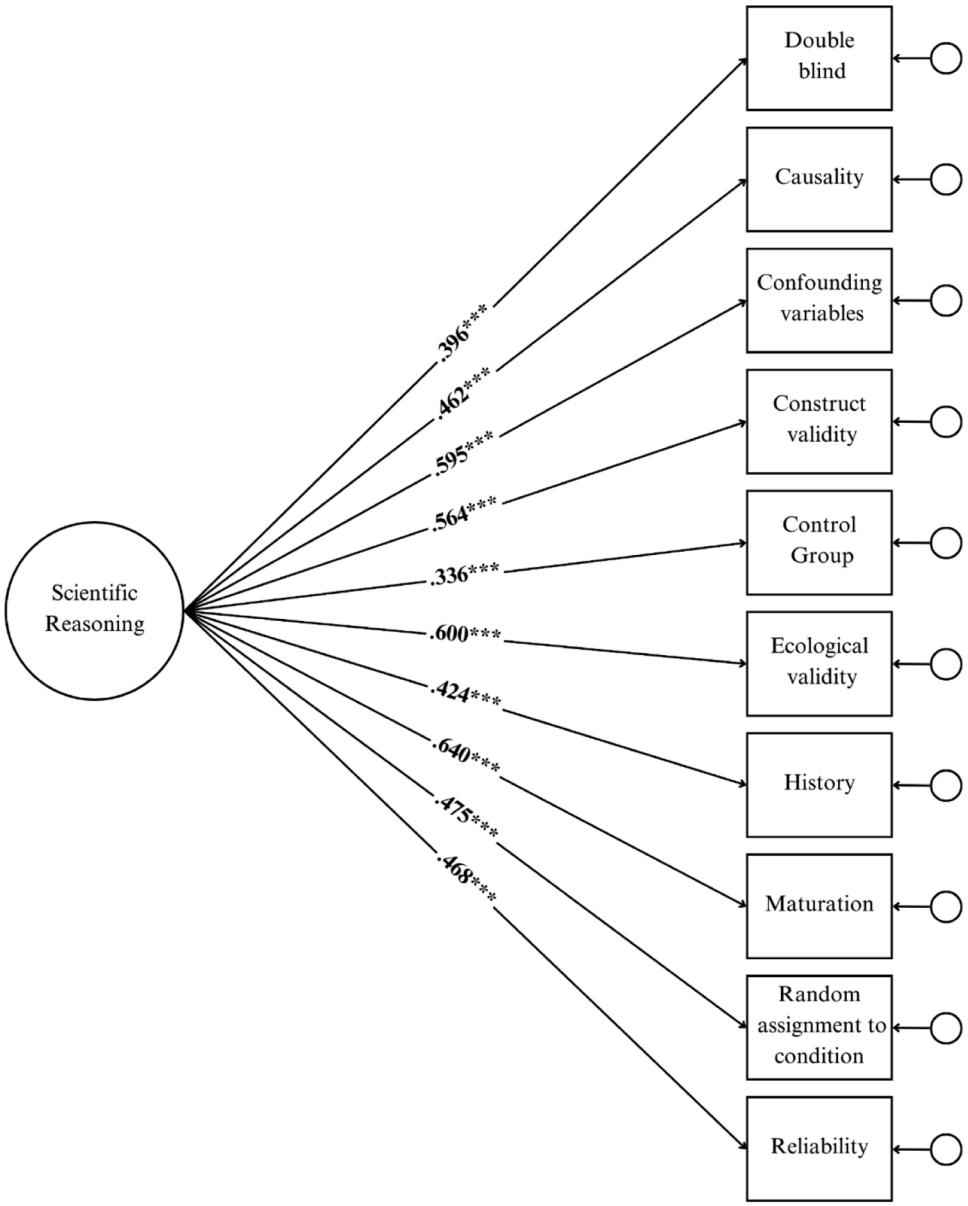


Fig. 2. Confirmed factorial structure of scientific reasoning scale (SRS).
Note. Numbers on the arrows indicate standardized factor loadings (range: 0.336 to 0.640). The items significantly load onto a single latent factor representing scientific reasoning (*** indicates $p < .001$).

achieved, as indicated by a significant difference between the strict and factor invariance models (Difftest $\chi^2 < 0.001$; $\Delta CFI = -.033$). In this case, participants with a university degree showed significantly higher factor means than those with only a high school diploma.

Convergent evidence. Convergent validity was assessed by correlating the SRS with measures of Cognitive Reflection Test (CRT-Long) and Probabilistic Reasoning Scale (PRS) on a sample of 896 participants. The standardized correlation coefficients indicated moderate to strong positive associations: the SRS correlated 0.523 ($p < .001$) with the CRT-Long and 0.578 ($p < .001$) with the PRS. In the model, correlations were also required between the two kinds of convergent measures (CRT-Long and PRS) in order to control for their covariance. The fit of this model was good [χ^2 (53) = 88.14, $p = .002$; RMSEA = 0.027 (0.017 0.037), $p = 1$; CFI = 0.973; WRMR = 0.909].

Criterion-related evidence. Criterion-related evidence was examined on a sample of 896 participants by testing a SEM model in which the score of SRS was related to different outcomes: climate change awareness (CCA), climate change beliefs (CCB) and the five factors of the Paranormal Health Beliefs Scale (parapsychological beliefs - PSIB, superstitious beliefs - SB, religious beliefs - RB, extraordinary events beliefs - EEB, and pseudo-scientific beliefs - MEDB). The fit of the model was good [χ^2 (98) = 123.85, $p = .040$); RMSEA = 0.017 (0.004 0.026), $p = 1$; CFI = 0.989; WRMR = 0.750]. Negative correlations were found between the SRS and all five paranormal health

Table 3
Measurement invariance of the SRS across different demographic and ideological conditions.

Invariance	χ^2	Df	p	RMSEA (CI)	p RMSEA ≤ 0.05	CFI	WRMR	DIFF TESTχ^2	Df	p	ΔCFI
Gender (438 male vs 453 female)											
Configural	85.40	70	.102	.02 (0.00 0.04)	1.00	.980	1.12				
Weak	91.03	79	.167	.02 (0.00 0.03)	1.00	.984	1.22	8.53	9	.482	.004
Strong	114.55	88	.030	.03 (0.01 0.04)	1.00	.965	1.37	25.83	9	.002	−0.019
freeing item 9	103.44	87	.110	.02 (0.00 0.04)	1.00	.978	1.30	13.44	8	.098	−0.006
Strict								10.15	10	.428	
Factor Variance	102.11	88	.144	.02 (0.00 0.03)	1.00	.981	1.31	0.46	1	.496	.003
Factor Mean	104.50	89	.125	.02 (0.00 0.03)	1.00	.979	1.34	2.81	1	.094	−0.002
Age (180 GenZ vs 315 GenY vs 401 GenX)											
Configural	130.90	105	.044	.03 (0.01 0.04)	.99	.968	1.39				
Weak	141.48	123	.122	.02 (0.00 0.04)	1.00	.977	1.54	16.26	18	.574	.009
Strong	163.21	141	.097	.02 (0.00 0.04)	1.00	.972	1.65	22.62	18	.206	−0.005
Strict								25.27	20	.191	
Factor Variance	187.91	143	.007	.03 (0.02 0.04)	.99	.944	1.83	11.39	2	.003	−0.028
Factor Mean	189.32	145	.008	.03 (0.02 0.04)	1.00	.945	1.85	2.83	2	.243	.001
Education level (536 up to the diploma vs 356 up to the PhD)											
Configural	82.91	70	.139	.02 (0.00 0.04)	1.00	.984	1.10				
Weak	90.48	79	.178	.02 (0.00 0.03)	1.00	.986	1.21	9.61	9	.383	.002
Strong	107.18	88	.081	.02 (0.00 0.04)	1.00	.976	1.32	18.14	9	.034	−0.010
Strict								19.52	10	.034	
Factor Variance	134.88	89	<0.001	.03 (0.02 0.05)	.99	.943	1.53	17.03	2	<0.001	−0.033
Employment status (134 students vs 588 workers vs 173 'other')											
Configural	126.79	105	.073	.03 (0.00 0.04)	1.00	.971					
Weak	136.39	123	.193	.02 (0.00 0.04)	1.00	.982	1.52	14.87	18	.671	.011
Strong	158.95	141	.143	.02 (0.00 0.04)	1.00	.976	1.64	24.10	18	.152	−0.006
Strict								25.09	20	.198	
Factor Variance	171.18	143	.054	.03 (0.00 0.04)	1.00	.963	1.74	7.52	2	.023	−0.013
Factor Mean	172.49	145	.059	.03 (0.00 0.04)	1.00	.964	1.76	2.49	2	.288	.001
Political orientation (333 right vs 294 center vs 269 left)											
Configural	135.35	105	.025	.03 (0.01 0.05)	.988	.960	1.42				
Weak	144.65	123	.089	.02 (0.00 0.04)	.999	.971	1.57	15.99	18	.594	.011
Strong	163.06	141	.099	.02 (0.00 0.04)	1.00	.971	1.66	18.55	18	.420	.000
Strict								23.27	20	.276	
Factor Variance	162.40	143	.128	.02 (0.00 0.04)	1.00	.974	1.68	2.00	2	.368	.003
Factor Mean	181.44	145	.022	.03 (0.01 0.04)	.998	.952	1.81	11.17	2	.004	−0.022
Religious affiliation (474 'yes' vs 423 'no')											
Configural	97.13	70	.018	.03 (0.01 0.04)	.996	.966	1.20				
Weak	111.28	79	.010	.03 (0.02 0.04)	.997	.959	1.36	14.80	9	.097	−0.007
Strong	117.89	88	.018	.03 (0.01 0.04)	.999	.962	1.40	6.65	9	.674	−0.003
Strict								12.29	10	.266	
Factor Variance	118.52	89	.020	.03 (0.01 0.04)	.999	.962	1.42	1.33	1	0,2483	.000
Factor Mean	132.00	90	.003	.03 (0.02 0.04)	.996	.947	1.52	7.00	1	0,0082	−0.015

belief factors: PSIB ($r = –.316$, $p < .001$), SB ($r = –.421$, $p < .001$), RB ($r = –.398$, $p < .001$), EEB ($r = –.406$, $p < .001$), and MEDB ($r = –.394$, $p < .001$). Standardized correlations indicated that the SRS was positively associated with CCA ($r = 0.118$, $p < .01$) and CCB ($r = 0.131$, $p < .01$).

4. Discussion

Using a sample of 897 Italian participants aged 18 and over, the present study aimed to provide validity evidence for the Italian adaptation of the SRS, through SEM, in line with the contemporary view of validity (Hubley & Zumbo, 2011). The results offer strong evidence of the instrument's psychometric robustness and cross-group applicability, addressing key gaps in the measurement of scientific reasoning within the Italian context.

4.1. Summary and interpretation of findings

Confirmatory factor analysis supported the unidimensional structure of the Italian SRS, with all ten retained items showing acceptable loadings (>0.30) after the removal of one item (response bias) due to a low loading. Although some loadings were modest (e.g., items related to 'Control group' and 'Double blind'), this is expected for broad constructs with diverse facets (Brown, 2015; Kline, 2023). Facets represent distinct aspects of a broad construct (e.g., blind/double-blind, causality, confounding variables, construct validity, control group, ecological validity, history, maturation, random assignment, reliability, and response bias) that together define its conceptual breadth. As Kline (2023) notes, constructs with multiple facets often yield modest loadings for some items, since each indicator taps a different facet while still contributing to the overarching latent dimension. Composite reliability was modest but

acceptable, likely reflecting the limited number of items (as reliability typically increases with scale length; Cortina et al., 2020; Graham, 2006). It also reflects the conceptual breadth of the scientific reasoning construct, which spans ten distinct content domains or facets (Little et al., 1999).

Measurement invariance analyses demonstrated the scale's robustness across gender, age, employment status, political orientation, and religious affiliation. These findings suggest that the SRS functions equivalently across different groups, enabling meaningful comparisons. 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.

In line with previous studies from the U.S., Israel, and Turkey (Drummond & Fischhoff, 2017; Golumbic et al., 2023; Muslu Kaygisiz et al., 2018), no gender differences were observed. However, in our sample, for strong invariance in the gender comparison, item 9 ('Some researchers want to verify if a psychoeducational intervention helps children improve their eating habits. The children participating in the study will be divided into an intervention group and a control group') needed to be freed: females were more likely than males to answer incorrectly despite equivalent latent trait levels. One possible explanation is that emotional, ethical, or motivational factors, such as stronger affective responses to child-related interventions, might differentially influence reasoning processes in this context (Golumbic et al., 2023; Shah et al., 2017). From a validity perspective (AERA, APA, & NCME, 2014; Zumbo & Hubley, 2017), this tentative finding suggests a line of inquiry for future research aimed at examining potential sources of construct-irrelevant variance. While this partial non-invariance does not undermine the validity of the scale as a whole, it highlights the value of further investigation to determine whether and how such factors may affect item functioning across gender groups.

No significant differences emerged across age, employment status, political orientation, or religious affiliation, indicating that the SRS operates equivalently for these groups. For these variables, however, our findings of invariance only partially align with earlier work, where results have been more mixed. Such discrepancies may reflect methodological differences; for instance, the use of factor means in the present study versus raw score comparisons in others, or cultural factors that influence how demographic variables are experienced across contexts (Moyser, 2002).

Education level emerged as the only demographic variable with significant effects on latent means: participants with a university degree scored higher on scientific reasoning, even though structural invariance held. This finding reinforces the link between higher education and advanced reasoning skills, as highlighted in prior research (Bao et al., 2022) and international frameworks such as PISA 2025, which promote scientific reasoning as a core civic competency. At the same time, it signals the risk of uneven skill distribution across the population, pointing to the need for broader access to educational opportunities that cultivate scientific reasoning.

Convergent validity was confirmed by significant positive correlations between the SRS and both the Cognitive Reflection Test and the Probabilistic Reasoning Scale, replicating patterns observed in the original development and validation work. Criterion-related validity was also supported, as higher scientific reasoning scores were associated with lower endorsement of paranormal and pseudoscientific beliefs, and with greater awareness of and attribution of climate change to human activity, although the correlations with climate change variables were small, again in line with previous findings (Drummond & Fischhoff; 2017).

4.2. *Methodological contributions and gaps addressed*

This study also offers several methodological strengths. First, the use of a quota-based sample of Italian adults, designed to approximate national representativeness through stratification by age, gender, and geographical region, increases the stability and generalizability of the results. Second, the validation process followed best practices for measurement transparency, as recommended by Flake and Fried (2020), by explicitly documenting scale selection, adaptations, and psychometric decisions. Third, the research embraced principles of open science, following the four-pronged framework proposed by Gai et al. (2025), which promotes transparent design, contextualized interpretation, responsible generalization, and inclusive evaluation practices. These elements reinforce the validity and replicability of our findings.

Moreover, this study addresses two significant gaps in existing literature. First, to date, no previous research has employed SEM to provide validity evidence for the SRS, nor has measurement invariance been tested; as a result, no empirical evidence has supported the assumption that the scale functions equivalently across different groups or cultural contexts. Second, no measure of scientific reasoning with established validity evidence is currently available for the Italian population. Providing validity evidence for psychological instruments across cultural contexts is essential not only for extending the cross-national applicability of a construct, but also for contributing additional support for validity, particularly in terms of generalizability across diverse administration settings. This approach is consistent with contemporary perspectives on validity (Hubley & Zumbo, 2011), which emphasize the importance of gathering evidence from multiple sources and contexts. Furthermore, identifying and defining constructs is a foundational step in the development of psychological theory, as empirical research relies on the availability of valid tools to measure such constructs. Construct validation, defined as the process of collecting evidence to confirm that an instrument measures what it claims to measure, is a difficult and necessary part of the research process (Cronbach & Meehl, 1955; Flake & Fried, 2020). By addressing both of these gaps, this study contributes to the international development of scientific reasoning as a measurable and culturally generalizable construct.

4.3. *Educational and social implications*

Beyond its psychometric validation, this work underscores the societal relevance of scientific reasoning as a core competency for navigating complex information, evaluating evidence, and resisting misinformation, skills that, as emphasized by Allchin et al. (2024), are increasingly critical in today's media and information landscape. The uneven distribution of these skills, particularly along educational lines, calls for the systematic inclusion of scientific reasoning instruction in curricula, starting from compulsory education.

A further consideration concerns cultural and educational characteristics specific to the Italian context. Italian schooling has traditionally emphasized disciplinary content knowledge over inquiry-based or metacognitive approaches, with limited systematic instruction in reasoning about evidence or uncertainty (Fasanelli et al., 2024). Moreover, national curricula tend to focus on scientific content rather than on the epistemic processes underlying science learning. Recent Eurobarometer surveys (Publications Office of the European Commission, 2021, 2025) consistently show that Italian adults display comparatively lower levels of trust in science and lower self-reported understanding of scientific information compared with the EU average. These findings point to a broader cultural climate in which scientific reasoning is often perceived as remote from everyday thinking. Such structural and cultural features may partly explain the association between higher education and stronger scientific reasoning observed in our data, as formal exposure to research methods and epistemological reflection usually occurs only in advanced stages of education. Strengthening inquiry-oriented teaching practices and explicitly incorporating reasoning about evidence into earlier levels of schooling could therefore represent a key educational priority.

The Italian SRS has potential applications in multiple domains: in education, it can help identify reasoning gaps, guide curriculum design, and evaluate interventions; in science communication and outreach, it can support efforts to monitor and promote evidence-based reasoning among the general public.

4.4. Limitations and future directions

Several limitations should be acknowledged. First, the SRS items represent typical laboratory research scenarios, and participants are required to evaluate an artificial context and select the correct answer. The scenario used may be less accessible to individuals without formal scientific training. As such, the ecological validity of the measure may be limited, and generalizations to real-world reasoning should be made with caution. Future studies could complement this measure with instruments grounded in everyday scenarios to assess informal or intuitive scientific reasoning (e.g., Golumbic et al., 2023).

Second, the change of response format represents an adaptation rather than a direct translation of the original SRS. Although the core content and intent of the items were preserved, this change may have influenced response dynamics and measurement properties. The adapted version improves ecological and linguistic fit in the Italian context, enhancing clarity and interpretability of responses. However, it may also limit the direct comparability of findings with studies using the original format.

Third, one item, assessing response bias, was removed due to insufficient factor saturation. While this exclusion improved the psychometric fit of the model, it also resulted in the loss of a meaningful aspect of the scientific reasoning construct. Future studies should explore alternative ways to assess response bias and incorporate it into the scale in a psychometrically robust manner.

Fourth, while confirmatory factor analysis confirmed a unidimensional structure, some items showed modest factor loadings, suggesting variability in how strongly each item contributes to the latent construct. This points to the potential benefit of refining specific items or developing parallel forms to enhance conceptual clarity and measurement precision. Future research could also employ alternative measurement approaches, such as item response theory (IRT) models, to examine more closely how individual items function in relation to the construct and to provide a more fine-grained understanding of the scale's internal structure.

Fifth, although the composite reliability is within an acceptable range for a brief psychological scale, it remains modest. This is likely due to the combination of brevity and conceptual breadth: each item reflects a distinct facet of scientific reasoning, which reduces inter-item correlations and thus internal consistency. This trade-off was intentional to ensure broad coverage of the construct. Nevertheless, future research might increase the number of items within each facet or develop facet-specific subscales to improve internal consistency without sacrificing construct representation. Moreover, the cross-sectional design limits the assessment of temporal stability, which could be addressed through longitudinal or test–retest studies to examine score consistency over time or in response to interventions.

Sixth, partial measurement invariance was observed for item 9 in gender comparisons, suggesting that emotional, ethical, or motivational factors may differentially affect responses. Future research should investigate the sources of this construct-irrelevant variance and consider refining or replacing the item to enhance measurement fairness. In addition, alternative measurement approaches such as item response theory (IRT) could be employed to examine item functioning in greater detail and to detect potential sources of differential item functioning (DIF) across groups.

Finally, in line with the contemporary view of validity (e.g., Hubley & Zumbo, 2011), it is important to recognize that construct validation is an ongoing process rather than a one-time achievement. Additional evidence, such as ecological validity from real-life contexts, or psychometric robustness using alternative models like Item Response Theory (IRT), could complement the findings presented here. Moreover, as the SRS now exists in at least four countries (Italy, the US, Turkey, and Israel), future cross-cultural studies should investigate measurement invariance and potential cultural differences in scientific reasoning.

5. Conclusion

This study provides the first validity evidence for score interpretations from the Scientific Reasoning Scale in the Italian adult population, delivering robust evidence of its psychometric soundness and cross-group comparability. By confirming a unidimensional structure, establishing measurement invariance across key demographic variables, and demonstrating convergent and criterion-related validity, we offer a reliable and culturally adapted tool for assessing scientific reasoning in Italy. The methodological rigor adopted, including the use of SEM, a quota-based sample approximating national representativeness, and transparency in reporting, also advances best practices in providing validity evidence for measurement instruments.

Theoretically, our findings strengthen the conceptualization of scientific reasoning as a broad yet coherent construct that is

measurable across cultural contexts. The scale's demonstrated invariance supports its use in cross-national research, fostering cumulative theoretical development in the psychology of reasoning. Practically, the Italian SRS, supported by validity evidence, offers an evidence-based instrument for educators, policymakers, and science communicators to diagnose reasoning gaps, inform targeted interventions, and monitor progress in fostering evidence-based thinking. The observed link between higher education and stronger reasoning skills underscores the importance of embedding scientific reasoning instruction early and systematically in formal education, as well as promoting lifelong learning opportunities to reduce skill disparities in the wider population.

By bridging a significant gap in the availability of measures with established validity evidence, this study not only extends the reach of the SRS but also contributes to the international effort to conceptualize, assess, and strengthen scientific reasoning, an essential competency for informed citizenship in an era of complex global challenges.

Ethics approval

The study protocol was approved by the Ethical Committee of the Department of Psychology of Università Cattolica del Sacro Cuore of Milan and performed in accordance with the ethical standards as laid down in the 1964 Declaration of Helsinki and its later amendments.

Consent to participate and share date

Informed consent and permission to share data on open science platforms, making responses publicly available for research, were obtained from all individual participants included in the study.

Data publicly available in a repository

To ensure the reproducibility of this study, in accordance with open science principles, we provide all methodological details, along with the scripts and data available on OSF. Specifically, the OSF page includes data, project preregistration, employed statistics, quotas for obtaining a representative sample, a codebook, and the analysis scripts.

- Study 1: https://osf.io/jk9dp/?view_only=ca1748d2cd474828bd78fddb7b97a4bc.
- Study 2: https://osf.io/6xrw7/?view_only=58b191ae5a224080929d83840baf7315.

However, we would like to specify that in our project preregistration, we stated that we would test a second version of the Scientific Reasoning Scale and that we intended to provide validity evidence for both scales using both classical test theory and item response theory. However, we have come to realize that such a project would not be feasible due to its complexity and scope. Consequently, we reserve the right to fulfill this intention in a future study that will utilize a different data collection.

Funding

No funding was received for conducting this study.

CRediT authorship contribution statement

Rossella Caliciuri: Writing – original draft, Visualization, Resources, Project administration, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Margherita Lanz:** Writing – review & editing, Validation, Supervision, Resources, Project administration, Methodology, Data curation, Conceptualization.

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.tsc.2025.102064.

References

Allchin, D., Bergstrom, C. T., & Osborne, J. (2024). Transforming science education in an age of misinformation. *Journal of College Science Teaching, 53*(1), 40–43. https://doi.org/10.1080/0047231X.2023.2292409

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education (Eds.). (2014). *Standards for educational and psychological testing.* American Educational Research Association.

Asparouhov, T., & Muthén, B. (2018, May). *SRMR in Mplus*.

Bao, L., Cai, T., Koenig, K., Fang, K., Han, J., Wang, J., … Wu, N. (2009). Learning and scientific reasoning. *Science, 323*(5914), 586–587. https://doi.org/10.1126/science.1167740

Bao, L., Koenig, K., Xiao, Y., Fritchman, J., Zhou, S., & Chen, C. (2022). Theoretical model and quantitative assessment of scientific thinking and reasoning. *Physical Review Physics Education Research, 18*(1), Article 010115. https://doi.org/10.1103/PhysRevPhysEducRes.18.010115

Bertolotti, M., Catellani, P., & Nelson, T. (2021). Framing messages on the economic impact of climate change policies: Effects on climate believers and climate skeptics. *Environmental Communication, 15*(6), 715–730. https://doi.org/10.1080/17524032.2021.1890175
Bontempo, D. E., & Hofer, S. M. (2007). Assessing factorial invariance in cross-sectional and longitudinal studies. In A. D. Ong, & M. H. M. van Dulmen (Eds.), *Oxford handbook of methods in positive psychology* (pp. 153–175). Oxford University Press. https://doi.org/10.1093/oso/9780195172188.003.0011.
Brown, T. A. (2015). *Confirmatory factor analysis for applied research* (2nd ed.). The Guilford Press.
Cheung, G. W., & Rensvold, R. B. (2002). Evaluating goodness-of-fit indexes for testing measurement invariance. Structural equation modeling. *A Multidisciplinary Journal, 9*(2), 233–255. https://doi.org/10.1207/S15328007SEM0902_5
Citizens' knowledge, perceptions, values and expectations of science – Report. (2021). Publications Office of the European Union. https://data.europa.eu/doi/10.2775/071577.
Cortina, J. M., Sheng, Z., Keener, S. K., Keeler, K. R., Grubb, L. K., Schmitt, N., Tonidandel, S., Summerville, K. M., Heggestad, E. D., & Banks, G. C. (2020). From alpha to omega and beyond! A look at the past, present, and (possible) future of psychometric soundness in the *journal of applied psychology*. *Journal of Applied Psychology, 105*(12), 1351–1381. https://doi.org/10.1037/apl0000815
Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin, 52*(4), 281–302. https://doi.org/10.1037/h0040957
de Oliveira Cardoso, N., de Lara Machado, W., Sorgente, A., & Guilherme, A. A. (2024). Cross-cultural adaptation and validation of the multidimensional subjective financial well-being scale in Brazilian adults. *Journal of Family and Economic Issues*, 1–21. https://doi.org/10.1007/s10834-024-09965-9
DeVellis, R. F., & Thorpe, C. T. (2021). *Scale development: Theory and applications*. Sage publications. https://doi.org/10.1111/peps.12499
Di Battista, S., Pivetti, M., & Berti, C. (2018). Moral foundations, political orientation and religiosity in Italy. *The Open Psychology Journal, 11*(1). https://doi.org/10.2174/1874350101811010046
Díaz, C., Dorner, B., Hussmann, H., & Strijbos, J. W. (2023). Conceptual review on scientific reasoning and scientific thinking. *Current Psychology, 42*(6), 4313–4325. https://doi.org/10.1007/s12144-021-01786-5
Dimitrov, D. M. (2010). Testing for factorial invariance in the context of construct validation. *Measurement and Evaluation in Counseling and Development, 43*(2), 121–149. https://doi.org/10.1177/0748175610373459
Donizzetti, A. R., & Petrillo, G. (2017). Validation of the paranormal health beliefs scale for adults. *Health Psychology Open, 4*(2), Article 2055102917748460. https://doi.org/10.1177/2055102917748460
Drummond, C., & Fischhoff, B. (2017). Development and validation of the scientific reasoning scale. *Journal of Behavioral Decision Making, 30*(1), 26–38. https://doi.org/10.1002/bdm.1906
Dunn, T. J., Baguley, T., & Brunsden, V. (2014). From alpha to omega: A practical solution to the pervasive problem of internal consistency estimation. *British Journal of Psychology, 105*(3), 399–412. https://doi.org/10.1111/bjop.12046
European citizens' knowledge and attitudes towards science and technology – Eurobarometer report. (2025). Publications Office of the European Union. https://data.europa.eu/doi/10.2777/6040908.
Eysenck, M. W., & Keane, M. T. (2020). Cognitive psychology: A student's handbook. *Psychology Press*. https://doi.org/10.4324/9781351058513
Fan, X., & Sivo, S. A. (2007). Sensitivity of fit indices to model misspecification and model types. *Multivariate Behavioral Research, 42*(3), 509–529. https://doi.org/10.1080/00273170701382864
Fasanelli, R., Piscitelli, A., & Di Lisio, M. (2024). La misurazione della fiducia nella scienza e negli scienziati: Adattamento italiano della scala in Science and Scientists. *Psicologia Della Salute: Quadrimestrale di Psicologia e Scienze Della Salute, 1*(2024), 125–139. https://doi.org/10.3280/PDS2024-001007
Fischer, F., Kollar, I., Ufer, S., Sodian, B., Hussmann, H., Pekrun, R., … Eberle, J. (2014). Scientific reasoning and argumentation: Advancing an interdisciplinary research agenda in education. *Frontline Learning Research, 2*(3), 28–45. https://doi.org/10.14786/flr.v2i2.96
Flake, J. K., & Fried, E. I. (2020). Measurement schmeasurement: Questionable measurement practices and how to avoid them. *Advances in Methods and Practices in Psychological Science, 3*(4), 456–465. https://doi.org/10.1177/2515245920952393
Frederick, S. (2005). Cognitive reflection and decision making. *Journal of Economic perspectives, 19*(4), 25–42. https://doi.org/10.1257/089533005775196732
Garnham, A., & Oakhill, J. (1994). *Thinking and reasoning*. Basil Blackwell.
Ghai, S., Thériault, R., Forscher, P., Shoda, Y., Syed, M., Puthillam, A., … Singh, L. (2025). A manifesto for a globally diverse, equitable, and inclusive open science. *Communications Psychology, 3*(1), 16. https://doi.org/10.1038/s44271-024-00179-1
Golumbic, Y. N., Dalyot, K., Barel-Ben David, Y., & Keller, M. (2023). Establishing an everyday scientific reasoning scale to learn how non-scientists reason with science. *Public Understanding of Science, 32*(1), 40–55. https://doi.org/10.1177/09636625221098539
Graham, J. M. (2006). Congeneric and (essentially) tau-equivalent estimates of score reliability: What they are and how to use them. *Educational and Psychological Measurement, 66*(6), 930–944. https://doi.org/10.1177/0013164406288165
Hu, L. T., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal, 6*(1), 1–55. https://doi.org/10.1080/10705519909540118
Hubley, A. M., & Zumbo, B. D. (2011). Validity and the consequences of test interpretation and use. *Social Indicators Research, 103*, 219–230. https://doi.org/10.1007/s11205-011-9843-4
IBM SPSS Statistics for Windows (Version 29.0) [Computer software]. *IBM Corp*, (2023).
Istituto Nazionale di Statistica. (2024). *Popolazione e indicatori demografici*. https://esploradati.istat.it/databrowser/#/it/dw/categories/IT1,POP,1.0/POP_POPULATION/DCIS_INDDEMOG1/IT1,22_293_DF_DCIS_INDDEMOG1_1,1.0.
Johnson, M. A., & Lawson, A. E. (1998). What are the relative effects of reasoning ability and prior knowledge on biology achievement in expository and inquiry classes? *Journal of Research in Science Teaching: The Official Journal of the National Association for Research in Science Teaching, 35*(1), 89–103. https://doi.org/10.1002/(SICI)1098-2736(199801)35:1<89::AIDTEA6>3.0.CO;2-J
Johnson-Laird, P. N., & Byrne, R. M. (1993). Mental models or formal rules? *Behavioral and Brain Sciences, 16*(2), 368–380. https://doi.org/10.1017/S0140525X0003065X
Kisiel, J., Rowe, S., Vartabedian, M. A., & Kopczak, C. (2012). Evidence for family engagement in scientific reasoning at interactive animal exhibits. *Science Education, 96*(6), 1047–1070. https://doi.org/10.1002/sce.21036
Kline, R. B. (2023). *Principles and practice of structural equation modeling*. Guilford publications.
Kowal, M. (2024). Translation practices in cross-cultural social research and guidelines for the most popular approach: Back-translation. *Anthropological Review, 87*(3), 19–32. https://doi.org/10.18778/1898-6773.87.3.02
Lai, K., & Green, S. B. (2016). The problem with having two watches: Assessment of fit when RMSEA and CFI dis-agree. *Multivariate Behavioral Research, 51*(2–3), 220–239. https://doi.org/10.1080/00273171.2015.1134306
Little, T. D. (1997). Mean and covariance structures (MACS) analyses of cross-cultural data: Practical and theoretical issues. *Multivariate Behavioral Research, 32*(1), 53–76. https://doi.org/10.1207/s15327906mbr3201_3
Little, T. D., Lindenberger, U., & Nesselroade, J. R. (1999). On selecting indicators for multivariate measurement and modeling with latent variables: When "good" indicators are bad and "bad" indicators are good. *Psychological Methods, 4*(2), 192–211. https://doi.org/10.1037/1082-989X.4.2.192
Marsh, H. W., Balla, J. R., & McDonald, R. P. (1988). Goodness-of-fit indexes in confirmatory factor analysis: The effect of sample size. *Psychological Bulletin, 103*(3), 391–410. https://doi.org/10.1037/0033-2909.103.3.391
Meade, A. W., Johnson, E. C., & Braddy, P. W. (2008). Power and sensitivity of alternative fit indices in tests of measurement invariance. *Journal of Applied Psychology, 93*(3), 568. https://doi.org/10.1037/0021-9010.93.3.568
Merenda, P. F. (1997). A guide to the proper use of factor analysis in the conduct and reporting of research: Pitfalls to avoid. *Measurement and Evaluation in Counseling and Development, 30*(3), 156–164. https://doi.org/10.1080/07481756.1997.12068936
Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist, 50*(9), 741.
Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd, pp. 13–103). Macmillan Publishing Co, Inc; American Council on Education.

Moyser, G. (Ed.). (2002). *Politics and religion in the modern world.* Routledge. https://doi.org/10.4324/9780203403778.
Muslu Kaygisiz, G., Gürkan, B., & Akbas, U. (2018). Adaptation of scientific reasoning scale into Turkish and examination of its psychometric properties. *Educational Sciences: Theory and Practice, 18*(3), 737–757. https://doi.org/10.12738/estp.2018.3.0175
Muthén, L. K., & Muthén, B. O. (1998-2017). *Mplus User's Guide.* Angeles, CA: Eighth Edition.Los.
Patton, M. Q. (2005). *Qualitative research.* New York, NY: John Wiley & Sons, Ltd. https://doi.org/10.1002/0470013192.bsa514
Peterson, R. A. (2000). A meta-analysis of variance accounted for and factor loadings in exploratory factor analysis. *Marketing Letters, 11*, 261–275. https://doi.org/10.1023/A:1008191211004
Primi, C., Morsanyi, K., Chiesi, F., Donati, M. A., & Hamilton, J. (2016). The development and testing of a new version of the cognitive reflection test applying item response theory (IRT). *Journal of Behavioral Decision Making, 29*(5), 453–469. https://doi.org/10.1002/bdm.1883
Primi, C., Morsanyi, K., Donati, M. A., Galli, S., & Chiesi, F. (2017). Measuring probabilistic reasoning: The construction of a new scale applying item response theory. *Journal of Behavioral Decision Making, 30*(4), 933–950. https://doi.org/10.1002/bdm.2011
Robba, M., Sorgente, A., & Iannello, P. (2024). search of socially responsible investors: A latent profile analysis. *Frontiers in Behavioral Economics, 3*, Article 1369261. https://doi.org/10.3389/frbhe.2024.1369261
Saris, W. E., Satorra, A., & van der Veld, W. M. (2009). Testing structural equation models or detection of misspe-cifications? Structural equation modeling. *A Multidisciplinary Journal, 16*(4), 561–582. https://doi.org/10.1080/10705510903203433
Shah, P., Michal, A., Ibrahim, A., Rhodes, R., & Rodriguez, F. (2017). What makes everyday scientific reasoning so challenging?. In *Psychology of learning and motivation, 66* pp. 251–299) Academic Press. https://doi.org/10.1016/bs.plm.2016.11.006
Sorgente, A., & Lanz, M. (2019). The multidimensional subjective financial well-being scale for emerging adults: Development and validation studies. *International Journal of Behavioral Development, 43*(5), 466–478. https://doi.org/10.1177/0165025419851859
Sorgente, A., Tagliabue, S., Andrade, C., Oliveira, J. E., Duan, W., & Lanz, M. (2021). Gender, age, and cross-cultural invariance of Brief Inventory of thriving among emerging adults. *Measurement and Evaluation in Counseling and Development, 54*(4), 251–266. https://doi.org/10.1080/07481756.2020.1827434
Sorgente, A., & Zumbo, B. (2025). The alphas and omegas of validity and reliability: Contemporary advances in evaluating and selecting instruments for quantitative research with emerging adults. In A. Sorgente, R. Vosylis, S. Claxton, & J. Schwab (Eds.), *Flourishing as a scholar: Research methods for the study of emerging adulthood* (pp. 92–112). Emerging Adulthood Series, Oxford University Press. https://doi.org/10.1093/oso/9780197677797.003.0007.
Toplak, M. E., West, R. F., & Stanovich, K. E. (2011). The Cognitive Reflection Test as a predictor of performance on heuristics-and-biases tasks. *Memory & Cognition, 39*(7), 1275–1289. https://doi.org/10.3758/s13421-011-0104-1
Widaman, K. F., Early, D. R., & Conger, R. D. (2014). Special populations. In T. D. Little (Ed.), *The oxford handbook of quantitative methods: Foundations* (pp. 55–81). Oxford University Press.
Willis, G. B. (2004). *Cognitive interviewing: A tool for improving questionnaire design.* sage publications.
Zimmerman, C. (2000). The development of scientific reasoning skills. *Developmental Review, 20*(1), 99–149. https://doi.org/10.1006/drev.1999.0497
Zumbo, B. D. (2005). Structural equation modeling and test validation. In B. Everitt, & D. C. Howell (Eds.), *Encyclopedia of statistics in behavioral science* (pp. 1951–1958). Chichester, UK: John Wiley & Sons Ltd. https://doi.org/10.1002/0470013192.bsa654.
Zumbo, B. D., & Hubley, A. M. (2017). *Understanding and investigating response processes in validation research.* Springer. https://doi.org/10.1007/978-3-319-56129-5