UNIVERSITÀ CATTOLICA DEL SACRO CUORE

Campus of Milano

Ph.D. Course in Psychology

XXXVIII cycle

S.S.D. PSIC-01/C

Scientific reasoning: theoretical foundations, measurement models, and validation studies

Supervisor:

Professor Margherita Lanz

Dissertation by:

Rossella Caliciuri

Id Number: 5215848

Academic Year 2024/2025

Finanziato dall'Unione europea NextGenerationEU

Ministero dell'Università e della Ricerca

Italiadomani PIANO NAZIONALE DI RIPRESA E RESILIENZA

Whatever exists at all exists in some amount.
To know it thoroughly involves knowing its quantity as well as its quality.
Education is concerned with changes in human beings; a change is a difference between two conditions; each of these conditions is known to us only by the products produced by it—things made, words spoken, acts performed, and the like.
To measure any of these products means to define its amount in some way so that competent persons will know how large it is, better than they would without measurement.
To measure a product well means so to define its amount that competent persons will know how large it is, with some precision, and that this knowledge may be conveniently recorded and used.
This is the general Credo of those who, in the last decade, have been busy trying to extend and improve measurements of educational products.

Edward L. Thorndike, *The Measurement of Intelligence* (1918)

Scientific reasoning: theoretical foundations, measurement models, and validation studies

Abstract: This dissertation examines scientific reasoning as both a psychological construct and a civic competence, addressing how it can be conceptually clarified, validly measured, and empirically validated. Despite its centrality to education and democratic participation, scientific reasoning has been variously defined – as cognitive strategy, epistemic stance, or social practice – resulting in theoretical fragmentation and measurement inconsistency. The present work aims to build a coherent validity argument for the measurement of SR, integrating conceptual analysis and empirical evidence within the contemporary unified framework of validity.

Following Loevinger's tripartite model of validation – substantive, structural, and external – the dissertation develops across five chapters. The first two chapters establish the theoretical and methodological foundations: Chapter 1 delineates scientific reasoning's definitions, components, and societal relevance, while Chapter 2 reframes validity as a property of the inferences and interpretations drawn from test scores, rather than of the test itself. It traces the evolution of validity theory toward an inference-based and cumulative framework, emphasizing that validation is an ongoing process of argumentation supported by multiple sources of evidence.

The empirical core comprises three studies. Study 1, a preregistered umbrella review, synthesizes systematic reviews and meta-analyses on scientific reasoning and adjacent constructs, mapping how they have been conceptualized, measured, and validated across decades of research. It reveals significant conceptual ambiguity and the lack of cumulative validity evidence. Study 2 presents the Italian adaptation and validation of the Scientific Reasoning Scale (SRS), gathering substantive, structural, and external validity evidence through confirmatory factor analyses, measurement invariance testing, and associations with theoretically relevant variables. Study 3 extends the structural validation by comparing measurement frameworks – Classical Test Theory and Item Response Theory – and by employing Latent Class Analysis as an exploratory, person-centered technique to investigate item performance.

Overall, these studies operationalize the logic of cumulative validation: from conceptual clarification to empirical testing, and from theoretical coherence to methodological refinement. The dissertation provides multiple sources of validity evidence for the Scientific Reasoning Scale and advances a transparent, cumulative approach to construct validation in psychological measurement.

Keywords: scientific reasoning, construct validity, contemporary view of validity, Classical Test Theory, Item Response Theory, Latent Class Analysis

Table of contents

Introduction

Every dissertation begins with a question, but often the question is born long before one enters academia. Mine began as an early feeling: that the scientific method is not merely a professional tool, but a deeply human way of making sense of the world. It lends credibility to our work as researchers, and speaks to something more intimate, something I have always felt: that within each of us lives a small, untrained scientist, endlessly testing our surroundings with curiosity and logic. I believe that this spirit never truly leaves us; it continues to guide even the most ordinary gestures of life.

I often think of my grandfather. He never had formal schooling, yet he possessed a kind of instinctive rigor. Whether he was tending the vineyard, adjusting the pruning of the vines, or caring for the wine resting in the cellar, he seemed to follow an unwritten protocol. I never knew where this knowledge came from, nor how he decided when to alter a gesture and when to repeat it faithfully, year after year. Perhaps it was trial and error, perhaps tradition, perhaps intuition disciplined by time. But what struck me most, looking back, was the logic behind it: an observation, a rule, an expected outcome, and the humility to repeat the cycle again. There was science in that rhythm—fragile, unrefined, but undeniably present. To me, it was an early reminder that scientific reasoning is not the privilege of laboratories, but a quiet inheritance of our humanity.

Scientific reasoning interests me, then, for at least three reasons. It is personal, because it was born from a personal curiosity about how we come to think scientifically, long before I began to study it formally. It is civic, because it represents a multifaceted competence at the core of education and democratic life. And it is professional, because it is the discipline I must practice every day as a researcher: designing studies, confronting uncertainty, tracing the links between theory and evidence, and accepting that knowledge advances through cumulative, transparent steps.

This personal, civic, and professional fascination has guided the work presented in this dissertation. From a conceptual perspective, it seeks to clarify what scientific reasoning is and how it relates to adjacent constructs. From a methodological standpoint, it aims to gather cumulative evidence of validity for its measurement, integrating qualitative and quantitative approaches across different studies. In doing so, the dissertation follows a coherent thread: to understand scientific reasoning not only as a psychological construct, but as a form of human inquiry that can, and should, be measured with rigor, transparency, and cultural sensitivity.

This dissertation is built upon two guiding foundations: one theoretical, concerning the construct of scientific reasoning, and one methodological, concerning the contemporary understanding of validity. These two dimensions – what is measured and how we establish the meaning of measurement – constitute the twin pillars of the work. Its overarching goal is to clarify the conceptual foundations of scientific reasoning and its relation to adjacent constructs, while providing cumulative validity evidence for its measurement.

The first part of the dissertation is devoted to two conceptual chapters. Chapter 1 concerns the construct of scientific reasoning – a competence whose definitions, domains, and boundaries remain the subject of lively

debate. It seeks to clarify this construct by examining its definitions, component skills, and outcomes, and by exploring its social and educational relevance as a form of epistemic competence. Chapter 2 addresses the contemporary framework of validity, tracing its evolution from a fragmented set of criteria to a unified, cumulative framework, and clarifying its implications for psychological measurement. Together, these two chapters provide the conceptual and methodological scaffolding on which the empirical studies are built.

After this theoretical groundwork, the dissertation moves to its empirical core, which consists of three studies. The first is an umbrella review that synthesizes systematic reviews and meta-analyses on scientific reasoning and related constructs, mapping how they have been conceptualized, measured, and validated across decades of research. The second presents the adaptation and validation of the Scientific Reasoning Scale for the Italian context, employing a quota-balanced sample and a multi-method design to collect validity evidence supporting the interpretation of test scores. The third offers a methodological comparison of the measurement models applied to the Scientific Reasoning Scale, combining Classical Test Theory and Item Response Theory with an exploratory Latent Class Analysis to examine item quality and response patterns.

The theoretical thread connecting these three studies is drawn from Jane Loevinger's seminal work *Objective Tests as Instruments of Psychological Theory* (1957). In that monograph, Loevinger proposed that validation is not a single event but a cumulative argument built across three interdependent components: substantive, structural, and external. The substantive component concerns the theoretical and conceptual justification of what the test is intended to measure, that is, the adequacy with which items represent the target construct; the structural component refers to the internal organization of the test, its dimensionality, homogeneity, and the coherence between observed patterns and theoretical expectations; finally, the external component encompasses relations between the test and external criteria, such as correlations with other measures, behavioral indicators, or theoretically relevant outcomes.

This tripartite framework offers not only a taxonomy of validity evidence but also a logical sequence for the empirical studies that follow. The umbrella review (chapter 3) applies this framework as an observational grid, examining how previous studies have conceptualized, measured, and validated scientific reasoning across the substantive, structural, and external phases. In doing so, it maps the state of the art, identifying both conceptual ambiguities and methodological gaps, particularly the scarcity of studies that have accumulated validity evidence according to the contemporary framework.

Building on these findings, the validation study (chapter 4) addresses these limitations by generating new empirical evidence within a clearly defined construct of scientific reasoning through the adaptation and validation of the Scientific Reasoning Scale in the Italian context. It encompasses all three phases of Loevinger's framework: the substantive phase, through the explicit theoretical definition and operationalization of the construct; the structural phase, through analyses of dimensionality and internal consistency; and the external phase, through the examination of relationships with theoretically relevant variables and outcomes. Together, these elements provide a comprehensive validation effort that directly responds to the conceptual and methodological gaps identified in the umbrella review and lay the groundwork

for further structural analyses, aimed at clarifying how individual items reflect and sustain the underlying construct.

Building on these results, the methodological comparison (chapter 5) represents a focused examination of the structural phase of Loevinger's framework. Unlike the previous study, it employs two measurement models, Classical Test Theory and Item Response Theory, to investigate the internal structure of the Italian Scientific Reasoning Scale and to assess the psychometric performance of its items. In addition, it introduces Latent Class Analysis as an exploratory, person-centered technique for gaining further insight into item quality. This approach allows for a deeper understanding of the internal functioning of the scale and refines the structural evidence gathered in the previous validation phase.

Considered collectively, these three studies embody the logic of cumulative validation that underlies the entire dissertation: from conceptual clarification to empirical testing, from theoretical coherence to methodological rigor, all aimed at advancing a unified understanding of *what scientific reasoning is* and *how it can be validly measured*. Each contributes a distinct layer of evidence – conceptual, structural, and external – to advance a more coherent understanding of scientific reasoning as both a psychological construct and a human capacity for inquiry. In this sense, validation is not treated as an endpoint but as an ongoing process, one that unfolds through the integration of theory, method, and empirical observation. The work as a whole aims to advance not only the measurement of scientific reasoning, but also our broader reflection on how psychological constructs acquire meaning through careful, cumulative inquiry.

This work is therefore both personal and scientific. It is personal because it stems from the intuition that science is the silent grammar of everyday life, not only the language of scholars. And it is scientific, because it seeks to contribute evidence, transparency, and rigor to the study of a construct that matters profoundly in today's world. At the intersection of these two paths lies the meaning of this dissertation: a recognition that the spark of reasoning scientifically belongs to all of us, and that our task, as researchers, is to illuminate it with the light of method.

Chapter 1 – Scientific Reasoning: Definitions, Societal Relevance, and Measurement

This chapter outlines how scientific reasoning has been defined, why it matters for individuals and societies, how it has been prioritized in educational frameworks, and how it has been measured. This sets the stage for the empirical work that follows.

1. Definitions and Domains of Scientific Reasoning

The study of scientific reasoning has its roots in developmental psychology and the cognitive sciences, where it was initially conceptualized as a set of reasoning strategies enabling individuals to generate, test, and revise knowledge claims. One of the most influential voices in this tradition is Deanna Kuhn (1989), who described scientific reasoning as a form of metacognitive inquiry that distinguishes scientific thinking from everyday reasoning. For Kuhn, the defining feature of scientific reasoning lies in the coordination between theories and evidence: individuals must be able to construct hypotheses, test them systematically, and then update their beliefs in light of data. This ability, however, is far from universal. Kuhn's studies demonstrated that many adolescents and even adults continue to rely on unsystematic reasoning strategies, failing to separate evidence from prior beliefs or to recognize the need for controlled experimentation.

In parallel, the work of Klahr and Dunbar (1988) introduced a more formalized cognitive model of how individuals engage in scientific discovery. Their *Scientific Discovery as Dual Search* (SDDS) framework portrays reasoning as a dynamic navigation between two interdependent spaces: the hypothesis space, where possible explanations are generated, and the experiment space, where empirical tests are designed. Scientific reasoning, in this view, emerges as the iterative cycle of proposing hypotheses, designing informative experiments, and using the results to refine or discard initial conjectures. This model highlighted the centrality of problem-solving processes and inspired a generation of research into how children and adults learn to reason scientifically in both laboratory and classroom settings.

Zimmerman (2000) added another layer by emphasizing the control-of-variables strategy (CVS) as a cornerstone of experimental reasoning. CVS refers to the ability to design or interpret experiments in which only one variable is manipulated at a time, holding others constant. Mastery of this strategy is often taken as an indicator of genuine scientific reasoning because it reflects an understanding of causal inference. Yet Zimmerman's review of the literature revealed that children's acquisition of CVS is gradual and fragile, requiring explicit instruction and scaffolding. Even adults, when faced with complex or ill-structured problems, frequently violate the principles of CVS, relying instead on intuitive heuristics.

Collectively, these contributions frame scientific reasoning as learnable rather than innate, and as requiring targeted, scaffolded educational experiences across development. This developmental perspective laid the empirical and conceptual foundations for later research, which would progressively broaden the construct

beyond its early experimental focus. While the early developmental and cognitive approaches laid the groundwork, subsequent decades saw the scope of scientific reasoning expand far beyond the laboratory. In the 2000s and 2010s, research in science education, argumentation studies, and decision sciences converged to redefine the construct as a multidimensional competence that is at once cognitive, epistemic, and social.

This shift resonates with Kuhn's (2001) influential account of epistemic development, which described how individuals progress from absolutist conceptions of knowledge (as certain and fixed), through multiplist views (knowledge as subjective opinions), toward evaluativist stances (knowledge as claims to be justified by evidence and argument). From this perspective, scientific reasoning is inseparable from epistemic cognition, as it requires not only the application of logical strategies but also an understanding of the status of knowledge claims.

A central contribution in this regard comes from Fischer and colleagues (2014), who proposed an interdisciplinary research agenda to integrate work on scientific reasoning and scientific argumentation. They defined scientific reasoning as a set of cognitive and epistemic processes aimed at generating and evaluating scientific knowledge, emphasizing that it involves not only experimental design but also argumentative practices, the critical evaluation of claims, and the ability to adopt an epistemic stance toward knowledge. In this view, reasoning scientifically means both "thinking like a scientist" and "participating in science as a social practice." The processes include identifying variables, generating hypotheses, drawing inferences, but also engaging in discourse, weighing counterarguments, and reflecting on the uncertainty of evidence.

This interdisciplinary perspective was further developed in Fischer et al. (2018), who highlighted the tension between domain-general and domain-specific conceptions. On the one hand, some aspects of scientific reasoning – such as the ability to recognize confounded designs or to coordinate theory and evidence – can be applied across contexts. On the other hand, reasoning cannot be fully understood without the conceptual knowledge specific to each domain. For example, interpreting an experiment in physics requires knowledge of physical principles, just as evaluating a study in psychology requires familiarity with constructs and measurement practices. Fischer and colleagues therefore argued that scientific reasoning emerges from the interplay between general epistemic processes and domain-specific conceptual knowledge, a position that resonates with current debates in science education.

The growing attention to epistemic processes also led researchers to highlight the overlap between scientific reasoning and epistemic cognition – that is, beliefs about the nature of knowledge and knowing. Bašnáková et al. (2021) demonstrated how individuals' epistemic beliefs shape their engagement in argumentation, influencing whether they treat claims as tentative and open to revision or as fixed facts. From this standpoint, scientific reasoning is not only about applying logical rules but also about understanding the epistemic status of knowledge claims. This shift broadens the domain of scientific reasoning to include epistemic humility, critical evaluation of sources, and sensitivity to uncertainty, qualities that are increasingly vital in a world saturated with competing information.

Díaz and colleagues (2021) synthesized this growing literature and noted the conceptual ambiguity that results from the proliferation of adjacent constructs such as scientific thinking, critical thinking, and scientific literacy. While each construct emphasizes slightly different aspects – cognitive strategies, evaluative dispositions, or application of science to real-life contexts – they share substantial overlap in practice. According to Díaz et al., scientific reasoning can be seen as a hub construct that links these related domains, capturing the processes that enable individuals to make sense of evidence, draw justified conclusions, and communicate arguments (2021). This implies that assessments should move beyond purely procedural skills to include epistemic and communicative practices.

Thus, from the mid-2010s onward, scientific reasoning has increasingly been conceptualized not as a narrow set of cognitive skills but as a broad, integrative competence that spans cognitive, epistemic, and social dimensions. This reconceptualization aligns with the needs of contemporary education systems, which seek to prepare students not only to master scientific facts but also to engage in evidence-based reasoning, to navigate uncertainty, and to participate in societal debates about science and technology.

In sum, current accounts converge on a hybrid view in which domain-general epistemic processes operate in tandem with domain-specific conceptual knowledge.

1.1 Contemporary Challenges and Everyday Contexts

While broader conceptualizations have enriched our understanding of scientific reasoning, they have also highlighted its fragility in practice. A growing body of research shows that both laypeople and students – even those with substantial science education – often struggle to apply scientific reasoning reliably in everyday contexts. Shah et al. (2017) argue that the core difficulty lies in theory – evidence coordination – that is, evaluating whether a body of evidence supports or contradicts a given claim. Their synthesis documents recurrent errors, including inferring causality from correlational data, overlooking threats to validity (e.g., selection bias), neglecting construct validity, and ignoring effect sizes or statistical uncertainty.

Building on this work, Michal and Shah (2024) demonstrate a specific bias in how people evaluate scientific findings in real-world decision-making contexts: when results are reported without numerical effect sizes, lay readers often assume practically large effects regardless of magnitude, a practical-significance bias that can lead to support for costly but minimally effective interventions.

These challenges have prompted conceptual and methodological shifts. Golumbic and colleagues (2023) introduce everyday scientific reasoning, defined as applying core methodological concepts (e.g., causality, reliability, random assignment) to scenarios outside the laboratory. They develop the Everyday Scientific Reasoning Scale to capture how citizens reason with evidence in authentic situations, showing that performance in daily-life scenarios is not equivalent to performance on structured, academic tasks.

Contemporary perspectives also emphasize the social and communicative dimensions of reasoning. Tabak and Dubovi (2023) argue that scientific reasoning must be understood within digital media ecologies, where

misinformation, persuasive anecdotes, and visuals compete with evidence-based communication, requiring navigation of conflicting claims and source credibility.

In synthesis, these lines of research portray scientific reasoning as both indispensable and precarious; indispensable for informed choices, judgments of claim trustworthiness, and civic participation, yet precarious because even well-educated individuals are prone to systematic biases and reasoning errors in complex, ill-structured settings.

In light of these challenges, contemporary accounts converge on scientific reasoning as a multifaceted, context-sensitive competence that spans cognitive strategies (developmental psychology), epistemic processes (argumentation research), and real-world applications (science communication). It functions as a bridge between scientific practices and everyday decision-making, and between disciplinary knowledge and civic engagement, a point that directly informs how the construct is measured and taught in the sections that follow.

2. Outcomes and Societal Relevance of Scientific Reasoning

This section clarifies why scientific reasoning matters beyond disciplinary contexts by tracing its impact across three planes. At the individual level, we examine how everyday and data-driven decision-making depends on evaluating evidence and causal claims. At the civic level, we consider its role in informed democratic participation and resistance to misinformation. At the educational level, we address the transfer of these skills across contexts and their links to critical thinking and argumentation. Collectively, these dimensions position scientific reasoning as a practical competence for participation in contemporary societies. As their importance has become increasingly evident, educational institutions and policy frameworks worldwide have progressively recognized scientific reasoning as a central goal of both education and citizenship.

One of the most compelling reasons why scientific reasoning has attracted sustained scholarly and educational interest is its relevance to everyday decision-making. Citizens constantly encounter scientific claims in domains such as health, nutrition, technology, and environmental sustainability. From reading headlines about the effectiveness of a new diet, to evaluating risks associated with vaccination, to interpreting reports about climate change, individuals must make choices that are both personally consequential and socially significant. The quality of these decisions depends heavily on their ability to interpret evidence, evaluate causal claims, and judge methodological soundness.

Yet research consistently demonstrates that people often fall short of these ideals. As Shah and colleagues (2017) observed, even educated adults frequently fail to notice methodological flaws in studies, misinterpret correlational evidence as causality, or ignore issues of effect size and statistical power (see also evidence on practical-significance biases; Michal & Shah, 2024). These biases are not merely academic curiosities: they have direct implications for people's health, financial stability, and well-being. A parent deciding whether to buy a costly nutritional supplement, a patient choosing between treatments, or a voter evaluating policy proposals all rely, implicitly or explicitly, on their scientific reasoning skills.

Furthermore, in contemporary societies, the relevance of scientific reasoning is amplified by the pervasiveness of data. The rapid growth of digital technologies, big data, and algorithmic decision-making has created what Bao et al. (2022) describe as a *world of data*. In such a world, citizens are expected not only to consume but also to critically evaluate data-driven claims. Here, scientific reasoning intersects with data literacy: the ability to interpret statistical information, recognize biases, and understand uncertainty. For example, interpreting a graph on COVID-19 infection rates or a report about the effectiveness of a new educational program requires more than factual knowledge. It demands the ability to question sampling procedures, to consider whether confounding variables might explain observed differences, and to evaluate whether the effect sizes reported are practically meaningful. Without these skills, individuals risk being misled by spurious correlations, sensationalist media coverage, or politically motivated distortions. Thus, scientific reasoning provides the epistemic backbone of data literacy, equipping individuals with the cognitive and methodological tools needed to make sense of the complex data environments that characterize the 21st century.

Scientific reasoning also plays a crucial role in civic life. Democratic societies increasingly require citizens to deliberate about issues with strong scientific components, ranging from vaccination policies to climate mitigation strategies. As Fischer et al. (2014) and Golumbic et al. (2023) note, the ability to evaluate evidence and to engage in informed argumentation is essential for meaningful participation in public debates. Without these competencies, citizens become more susceptible to misinformation, pseudoscience, and conspiracy narratives that proliferate in digital media ecosystems.

Survey data from the Eurobarometer (2021) illustrate this challenge. While trust in science remains high across Europe, significant misconceptions and cross-national disparities persist. On several indicators of scientific literacy and engagement, Italy performs below the EU average, highlighting how gaps in reasoning competence may hinder citizens' capacity for informed democratic participation. In this sense, scientific reasoning is not only a cognitive skill but also a civic necessity, one that sustains reasoned deliberation and strengthens democratic resilience.

Despite its recognized importance, scientific reasoning is often difficult to transfer beyond the classroom. Students may learn to apply the control-of-variables strategy in laboratory settings yet fail to use the same logic when evaluating media claims about nutrition or policy. This limited transfer underscores the need for educational approaches that cultivate reasoning in authentic contexts rather than through decontextualized exercises.

Recent work targets the problem of transfer by embedding tasks in realistic scenarios. Golumbic et al. (2023) developed the Everyday Scientific Reasoning Scale, contextualizing core methodological concepts in domains such as health and nutrition. Performance on these everyday tasks predicts the application of scientific information in real decisions, indicating that measuring reasoning in authentic contexts can enhance transfer beyond classroom-style tasks.

Scientific reasoning is also deeply intertwined with critical thinking and argumentation. As Bašnáková et al. (2021) show, these constructs converge in fostering evaluativist epistemic stances – viewing knowledge claims as tentative, evidence-based, and revisable (Kuhn, 2001). This stance enables individuals to engage in reflective judgment and resist misinformation. From this perspective, scientific reasoning is not only about evaluating evidence but also about adopting an epistemic stance that is open to revision in light of new data. This connection to critical thinking positions scientific reasoning as a crosscutting competence that supports resilience against misinformation, enhances problem-solving in professional contexts, and fosters lifelong learning. In an era where citizens are bombarded with competing claims and persuasive narratives, the ability to reason scientifically serves as both a protective factor and a transformative skill.

The growing awareness of scientific reasoning's importance at the individual, civic, and educational levels has naturally translated into institutional and policy emphasis. Educational systems across the world have progressively reframed science learning around reasoning, inquiry, and the evaluation of evidence, embedding these dimensions into international benchmarks and policy agendas. This evolution has culminated in the explicit recognition of scientific reasoning as a central goal of contemporary education.

The promotion of scientific reasoning as an educational goal can be traced back to the mid-20th century, when science curricula in the United States began shifting from rote memorization toward inquiry-based approaches. A milestone was the publication of the National Science Education Standards (NRC, 1996), which articulated a vision of science learning centered not only on mastering facts but on engaging in inquiry as a multifaceted activity – observing, asking questions, planning investigations, analyzing data, and communicating results. This vision marked a move away from viewing science as a static body of knowledge and reframed it as a dynamic process of reasoning. The emphasis on students' active engagement reflected broader reforms inspired by constructivist theories of learning. The standards argued that scientific reasoning should be practiced rather than merely taught, cultivating habits of mind characteristic of scientists. A decade later, the Framework for K–12 Science Education (NRC, 2012) advanced this vision by specifying scientific practices—eight crosscutting competencies that include asking questions, developing and using models, planning and carrying out investigations, analyzing and interpreting data, using mathematics and computational thinking, constructing explanations, engaging in argument from evidence, and obtaining, evaluating, and communicating information. Building on the Framework, the Next Generation Science Standards (NRC, 2013), developed by U.S. states, translated these practices into performance expectations and a three-dimensional model of learning that integrates practices, crosscutting concepts, and disciplinary core ideas. As noted by Rönnebeck et al. (2016), integrating reasoning with core disciplinary ideas represented a paradigmatic shift: reasoning became a central pillar of scientific literacy rather than an add-on to content.

In parallel, across the Atlantic, similar reforms were unfolding. In England, the launch of the *Twenty First Century Science* program signaled a broad reorientation of science education. This curriculum introduced the notion of "Ideas about Science", emphasizing understanding of the nature of scientific evidence, the role of uncertainty, and the evaluation of risk (Millar, 2006). The explicit aim was to prepare all students – not only

future scientists – with the capacity to evaluate claims in everyday contexts, from health advice to environmental debates. This program reflected a broader European trend toward equipping citizens with functional scientific literacy rather than simply deep disciplinary expertise.

At the European policy level, the European Commission and the Directorate-General for Research and Innovation (2015) advanced the agenda of *Science Education for Responsible Citizenship*. This report emphasized the importance of engaging students in inquiry-based, participatory approaches, including problem-solving, collaborative learning, and hands-on experiments. The rationale was that such pedagogies not only deepen conceptual understanding but also connect science to students' daily lives, empowering them to participate actively in democratic decision-making. In this framing, scientific reasoning was positioned as both a pedagogical goal and a civic competence.

Finally, the most influential global driver of attention to scientific reasoning has been the *Programme for International Student Assessment* (PISA), administered by the OECD. Since its inception in 2000, PISA has assessed 15-year-olds' scientific literacy, but the definition of literacy has progressively incorporated competencies directly tied to reasoning (2000, 2016). The 2023 framework identified three core competencies: (a) explaining phenomena scientifically, (b) evaluating and designing scientific inquiry, and (c) interpreting data and evidence (OECD, 2023). Each of these competencies explicitly requires students to engage in reasoning with evidence rather than merely recalling content knowledge. PISA's international reach has amplified the importance of reasoning, making it a policy priority worldwide. Nations compare their performance and adjust curricula accordingly, often with the goal of strengthening the reasoning competencies that underlie success on PISA assessments. This benchmarking function has reinforced the view that scientific reasoning is not only an academic skill but a strategic resource in a knowledge-based global economy.

Educational frameworks are not the only arena where scientific reasoning has gained prominence. Surveys of public opinion, such as the Eurobarometer 2021, reveal substantial variation in how European citizens perceive and engage with science. The survey documented high levels of trust in scientific institutions overall, but also persistent misconceptions and gaps in understanding. Notably, differences emerge between countries: citizens in Northern Europe tend to report stronger engagement with scientific issues, while countries such as Italy show lower levels of scientific literacy and confidence. These disparities suggest that educational systems play a crucial role in shaping not only knowledge but also reasoning dispositions and trust in science.

The link between education and public engagement is underscored by initiatives aimed at fostering citizen science and participatory engagement with research. These movements rest on the assumption that citizens must be able to reason scientifically – at least to some degree – to contribute meaningfully to collaborative projects and to evaluate the societal implications of emerging technologies. Thus, scientific reasoning is framed not only as a school outcome but as a lifelong competence necessary for informed citizenship.

Emerging trends point toward an even broader framing of scientific reasoning. As digital technologies transform how science is communicated and debated, reasoning is increasingly linked to media literacy and

digital literacy. Students are expected not only to evaluate controlled experiments but also to navigate online claims, social media debates, and algorithmically curated information streams (Tabak & Dubovi, 2023). This signals a new frontier in the educational emphasis on reasoning: preparing citizens not only to "think like scientists" but also to reason critically about science in complex, data-rich, and digitally mediated environments. This extends SR into digitally mediated environments, reinforcing its lifelong, civic relevance.

3. Measurement of Scientific Reasoning

The measurement of scientific reasoning presents a unique set of challenges. Unlike constructs with a relatively narrow theoretical core, scientific reasoning is multifaceted, contested, and context-dependent. As discussed in the preceding sections, researchers have conceptualized it variously as a domain-general cognitive strategy, a set of domain-specific practices, an epistemic stance, or a civic competence. Each of these conceptualizations implies different ways of operationalizing the construct, which explains the heterogeneity of measurement approaches that have emerged over the past four decades.

3.1 From Developmental Assessments to Epistemic Approaches

The earliest systematic attempts to measure scientific reasoning were embedded in developmental psychology. Researchers in the 1980s and 1990s focused on assessing children's ability to apply the CVS in experimental contexts (e.g., Chen & Klahr, 1999; Klahr & Dunbar, 1988; Zimmerman, 2000). Typical tasks involved presenting children with apparatuses such as ramps, pendulums, or mixing bottles and asking them to design or interpret experiments in which only one variable was manipulated while others were held constant (Chen & Klahr, 1999; Zimmerman, 2000). Performance on these tasks was taken as an indicator of whether children could reason about causal relations in a scientific way.

These developmental assessments provided invaluable insights into the cognitive underpinnings of scientific reasoning, documenting its progression from unsystematic to systematic experimentation. Yet their scope was narrow, capturing only procedural experimentation skills. They were often highly scaffolded and conducted in laboratory settings, raising questions about ecological validity and transferability. Moreover, they primarily targeted children or adolescents, leaving unexamined how reasoning skills manifest in older students or adults.

By the early 2000s, researchers began expanding the construct beyond experimental control. The focus shifted toward broader cognitive and epistemic dimensions, such as argumentation, evidence evaluation, and epistemic cognition (Kuhn, 2001; Fischer et al., 2014). New instruments included open-ended interviews and written justifications, as well as multiple-choice or scenario-based assessments requiring participants to evaluate claims, weigh evidence, or judge alternative explanations. This transition reflected an epistemological broadening – from viewing scientific reasoning as procedural problem solving to conceiving it as a socio-cognitive practice embedded in discourse and context.

These approaches highlighted that scientific reasoning is not just about manipulating variables but also about engaging with evidence in argumentative contexts. They reflected the growing consensus that reasoning must

be assessed not only as an individual cognitive process but also as a socially and discursively embedded practice.

3.2. Mapping the Landscape of Measurement Instruments

A systematic review by Opitz, Heene, and Fischer (2017) provided the most comprehensive overview to date of SR instruments. Analyzing 38 tests published between 1973 and 2013, the authors identified major trends in conceptualization, skill coverage, test formats, and psychometric quality. Building on the interdisciplinary framework by Fischer et al. (2014), they categorized scientific reasoning into eight core skills: problem identification, questioning, hypothesis generation, evidence generation, evidence evaluation, drawing conclusions, communication/scrutiny, and construction/redesign of artefacts. The review showed that most instruments capture only a limited subset of these skills, focusing primarily on hypothesis generation, evidence generation, evidence evaluation, and drawing conclusions. Early instruments such as the *Test of Logical Thinking* (TOLT; Tobin & Capie, 1981) and the *Classroom Test of Scientific Reasoning* (Lawson, 1978) were grounded in Piaget's (1958) stage theory of formal operational reasoning, treating scientific reasoning as a unidimensional cognitive ability. Their multiple-choice items typically asked students to predict experimental outcomes or identify controlled variables, thereby operationalizing scientific reasoning as mastery of logical control-of-variables strategies.

From the 2000s onward, a new wave of assessments reflected the shift toward multidimensional and process-oriented models, consistent with Klahr and Dunbar's (1988) *Scientific Discovery as Dual Search* (SDDS) framework. For instance, the *Experimenting as Problem Solving test* (Hammann et al., 2008) and the *Abilities in Scientific Inquiry scale* (Nowak et al., 2013) measured coordinated subskills – such as generating hypotheses, designing experiments, and evaluating data – within domain-specific contexts (biology and chemistry). These instruments typically employed scenario-based or open-ended tasks requiring students to plan and justify investigations, revealing a more authentic engagement with scientific reasoning processes. Similarly, the *Evidence-Based Reasoning Assessment System* (EBRAS; Brown et al., 2010) applied a rubric-based evaluation of students' written justifications, integrating the epistemic evaluation of evidence with argument quality. In parallel, large-scale educational assessments such as PISA adopted hybrid designs that combined multiple-choice and constructed-response items to capture the use of scientific evidence and reasoning in realistic contexts. In the 2010s, even more recent formats, such as computer-simulated environments (e.g., *Detector – Inquiry Intelligent Tutoring System*; Gobert et al., 2013), automatically analyze students' virtual experiments to assess how they generate and test hypotheses. This methodological evolution mirrors the educational transition from measuring factual knowledge to evaluating process-oriented and performance-based competencies.

Despite these advances, Opitz et al. (2017) highlighted persistent psychometric weaknesses. Fewer than half of the instruments reported reliability coefficients, and only a minority provided construct-validity evidence or model-based tests of dimensionality. Factor-analytic results were mixed: while some tests yielded single-

factor solutions consistent with a general scientific-reasoning ability (e.g., Test of Scientific Literacy Skills, TOSLS; Gormally et al., 2012), others supported multifactor structures reflecting differentiated reasoning components (e.g., Abilities in Scientific Inquiry, Nowak et al., 2013; Competence Scale for Learning Science, Chang et al., 2011). At a methodological level, these findings reflect a broader limitation – namely, the absence of systematic validation within the contemporary unified framework of validity (AERA, APA, & NCME, 2014; Messick, 1989). No instruments have accumulated evidence across substantive, structural, and external dimensions of validity, as recommended by the Standards for Educational and Psychological Testing. Most studies report isolated psychometric indices (e.g., reliability or factorial structure) rather than developing a cumulative validity argument that justifies the interpretation and use of scores. This gap makes it difficult to evaluate whether the inferences drawn from existing measures are warranted or whether their observed relations with other constructs truly support the intended interpretations.

The review underscored the limited generalizability of existing instruments, both in terms of target populations and disciplinary scope. The majority were designed for secondary or undergraduate students and anchored in specific domains such as biology, chemistry, or physics, reflecting a historical shift from domain-general assessments rooted in Piagetian reasoning to domain-specific operationalizations inspired by inquiry and problem-solving models. Only a few measures have been validated for general adult or professional populations, leaving significant gaps in our understanding of how adults outside formal education settings reason about scientific problems. This issue is particularly concerning given that scientific reasoning is increasingly recognized as a key civic competence, essential for navigating public debates, evaluating evidence, and making informed decisions in complex, information-rich societies.

The review by Opitz et al. (2017) covers instruments developed only up to 2013, leaving the post-2010s landscape largely unmapped. Given the proliferation of new conceptual frameworks and digital assessment tools in the past decade, further research is needed to obtain a comprehensive and updated overview of existing instruments. Recent conceptual reviews (e.g., Díaz et al., 2021) corroborate this heterogeneity and highlight persistent problems of operationalization and limited psychometric rigor—often reporting weak reliability, scarce construct validity, and inconsistent theoretical grounding. As Díaz et al. (2021) and Bašnáková et al. (2021) note, a core challenge lies in the lack of conceptual coherence across studies: when scientific reasoning is equated with critical thinking, assessments emphasize logical analysis and counterargumentation; when framed as scientific literacy, tasks focus on everyday or civic applications; and when defined as epistemic cognition, measures instead assess beliefs about knowledge and justification. This theoretical fragmentation leads to divergent operationalizations and undermines comparability, raising the question of whether existing instruments are indeed measuring the same underlying construct.

These conceptual, methodological, and psychometric challenges highlight the need for an instrument that articulates scientific reasoning in a precise and theoretically integrated way. The Scientific Reasoning Scale (SRS; Drummond & Fischhoff, 2017) responds to this need by providing a transparent measurement approach for evaluating how individuals appraise scientific evidence. Its clarity of construct definition represents a

crucial starting point for accumulating cumulative validity evidence, while its design for the general population extends the investigation of scientific reasoning beyond academic contexts toward its broader civic dimension.

3.3 The Scientific Reasoning Scale (SRS)

Building on the need for a clear and theoretically grounded operationalization outlined above, the Scientific Reasoning Scale (SRS; Drummond & Fischhoff, 2017) was developed to move beyond factual "scientific literacy" items and narrow developmental tasks, operationalizing reasoning as the ability to evaluate the quality of scientific findings. Its conceptual foundation integrates three traditions: the philosophy and methodology of science (which define normative criteria for sound research), the tradition of public understanding of science (which highlights what citizens already know about science and its methods), and cognitive developmental psychology (which examines how individuals acquire the ability to think scientifically).

Items present brief scenarios followed by true/false statements targeting core methodological concepts. The final version of the SRS includes 11 core methodological facets: blind/double blind, causality, confounding variables, construct validity, control group, ecological validity, history effects, maturation effects, random assignment to condition, reliability, and response bias. Each of these facets reflects a potential vulnerability in empirical research and, taken together, they capture the skills required to critically evaluate scientific findings.

In the original U.S. validation study, psychometric analyses indicated that the items loaded on a single latent factor, supporting the interpretation of the SRS as a unidimensional construct. Within that sample, the SRS demonstrated adequate reliability and strong construct validity, showing expected correlations with numeracy, cognitive reflection, and educational level. Importantly, the authors also reported incremental predictive validity over traditional measures of scientific literacy: higher SRS scores predicted not only better performance on comprehension tasks (e.g., interpreting drug facts boxes) but also beliefs more aligned with scientific consensus on topics such as vaccination, evolution, and genetically modified foods. These findings represent sample- and context-specific evidence for the intended score interpretations in the original U.S. study, which must be further accumulated and re-examined across populations and settings.

The present work focuses on the SRS as its primary measure of scientific reasoning for different reasons. First, it offers a clear and explicit operationalization of the construct, directly linking reasoning to the evaluation of evidence quality. Second, although it is not the only possible conceptualization of scientific reasoning, the SRS is distinctive in integrating three major research perspectives into a coherent framework, thereby offering a transdisciplinary operationalization that embodies the logic of the scientific method itself. Third, unlike many earlier instruments developed for students or disciplinary subfields, the SRS is explicitly designed for the general population, allowing the investigation of scientific reasoning as a civic competence rather than a purely academic one. This perspective aligns closely with a civic and applied view of scientific reasoning—one that moves beyond laboratory settings or classroom problem-solving, toward examining how individuals appraise the quality of empirical findings, detect methodological flaws, and judge the credibility of evidence presented

in real-world scenarios. For these reasons, the SRS represents a particularly relevant and timely framework for examining how people understand and use scientific reasoning as a cornerstone of evidence-based citizenship.

4. Concluding Remarks

The trajectory traced in this chapter highlights both the richness and the challenges of studying scientific reasoning. From its early conceptualizations in developmental psychology to contemporary interdisciplinary perspectives, scientific reasoning has been recognized as a multifaceted competence involving the coordination of theory and evidence, the application of domain-specific knowledge, and the ability to engage in critical argumentation and evidence-based decision-making. Its outcomes are evident not only in educational achievement but also in citizens' capacity to navigate a data-saturated world and to participate in democratic societies where science plays a central role. Accordingly, national and international educational frameworks have increasingly emphasized scientific reasoning as a key outcome of science education.

At the same time, the measurement of scientific reasoning remains a complex task. Existing instruments vary widely in scope and rigor, and few have undergone systematic validation across cultural contexts. The development of the Scientific Reasoning Scale (Drummond & Fischhoff, 2017) represents an important advance, yet to date it has not been validated within the contemporary framework of validity as articulated in the *Standards for Educational and Psychological Testing* (AERA, APA, & NCME, 2014). In particular, no studies have provided cumulative evidence across substantive, structural, and external dimensions of validity.

In the face of the conceptual complexity and fragmentation that characterize the study of scientific reasoning, a fundamental decision must be made: where to begin in bringing conceptual and methodological order to this construct. Scientific reasoning unfolds at multiple, interrelated levels – individual cognitive processes, everyday reasoning in daily contexts, and collective or social reasoning in public discourse. Ideally, a comprehensive account should integrate all these dimensions. Yet, before such integration is possible, a necessary first step is to ensure that the construct itself is clearly defined and measured with rigor. This need for conceptual clarity and measurement precision provides the rationale for the present dissertation. The first step is to clarify the construct itself, through a preregistered umbrella review that synthesizes how scientific reasoning has been conceptualized, measured, and validated across prior literature. Building on this foundation, the subsequent studies focus on the measurement of the Scientific Reasoning Scale in Italy: one study examines its internal structure and its relations to external constructs, and also assesses the generalizability of score interpretations; another integrates Classical Test Theory, Item Response Theory, and Latent Class Analysis to probe item quality and examine the complementarity of different psychometric perspectives.

This sequence of studies reflects the view that validation is not separate from conceptual work but deeply intertwined with it, unfolding as an ongoing, context-bound process of gathering multiple sources of evidence to support meaningful interpretations of test scores. Accordingly, the dissertation adopts a validity-as-argument perspective, in which each study adds a piece to a cumulative body of validity evidence.

The next chapter develops this framework of validity in greater depth, outlining its evolution, dimensions, and implications, and providing the methodological foundation for the empirical studies that follow.

Bibliography

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

Bašnáková, J., Čavojová, V., & Šrol, J. (2021). Does concrete content help people to reason scientifically? Adaptation of Scientific reasoning scale. *Science & Education*, *30*(4), 809-826. https://doi.org/10.1007/s11191-021-00207-0

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), 010115. https://doi.org/10.1103/PhysRevPhysEducRes.18.010115

Brown, N. J. S., Nagashima, S. O., Fu, A., Timms, M., & Wilson, M. (2010). A framework for analyzing scientific reasoning in assessments. *Educational Assessment, 15*(3-4), 142–174. https://doi.org/10.1080/10627197.2010.530562

Chang, H. P., Chen, C. C., Guo, G. J., Cheng, Y. J., Lin, C. Y., & Jen, T. H. (2011). The development of a competence scale for learning science: Inquiry and communication. *International Journal of Science and Mathematics Education*, *9*(5), 1213-1233. https://doi.org/10.1007/s10763-010-9256-x

Chen, Z., & Klahr, D. (1999). All other things being equal: Acquisition and transfer of the control of variables strategy. *Child Development, 70*(5), 1098–1120. https://doi.org/10.1111/1467-8624.00081

Díaz, C., Dorner, B., Hussmann, H., & Strijbos, J.-W. (2021). Conceptual review on scientific reasoning and scientific thinking. *Current Psychology: A Journal for Diverse Perspectives on Diverse Psychological Issues.* Advance online publication. https://doi.org/10.1007/s12144-021-01786-5

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

European Commission, Directorate-General for Research and Innovation. (2015). *Science education for responsible citizenship*. Publications Office of the European Union. https://doi.org/10.2777/12626

Citizens' knowledge, perceptions, values and expectations of science – Report, Publications Office of the European Union, 2021. https://data.europa.eu/doi/10.2775/071577

Fischer, F., Kollar, I., Ufer, S., Sodian, B., Hussmann, H., Pekrun, R., Neuhaus, B., Dorner, B., Pankofer, S., Fischer, M., Strijbos, J.-W., Heene, M., & 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

Fischer, F., Chinn, C. A., Engelmann, K., & Osborne, J. (Eds.). (2018). *Scientific reasoning and argumentation: The roles of domain-specific and domain-general knowledge*. Routledge. https://doi.org/10.4324/9780203731826

Gobert, J. D., Sao Pedro, M., Raziuddin, J., & Baker, R. S. (2013). From log files to assessment metrics: Measuring students' science inquiry skills using educational data mining. *Journal of the Learning Sciences*, *22*(4), 521-563. https://doi.org/10.1080/10508406.2013.837391

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

Gormally, C., Brickman, P., & Lutz, M. (2012). Developing a test of scientific literacy skills (TOSLS): Measuring undergraduates' evaluation of scientific information and arguments. *CBE—Life Sciences Education*, *11*(4), 364-377. https://doi.org/10.1187/cbe.12-03-0026

Hammann, M., Phan, T. T. H., Ehmer, M., & Grimm, T. (2008). Assessing pupils' skills in experimentation. *Journal of Biological Education*, *42*(2), 66-72. https://doi.org/10.1080/00219266.2008.9656113

Klahr, D., & Dunbar, K. (1988). Dual space search during scientific reasoning. *Cognitive science*, *12*(1), 1-48.

Kuhn, D. (1989). Children and adults as intuitive scientists. *Psychological review*, *96*(4), 674.

Kuhn, D. (2001). How Do People Know? *Psychological Science, 12*(1), 1–7. https://doi.org/10.1111/1467-9280.00302

Lawson, A. E. (1978). The development and validation of a classroom test of formal reasoning. *Journal of Research in Science Teaching, 15*, 11–24. http://doi.org/10.1002/tea.3660150103

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13–103). American Council on Education & Macmillan.

Millar, R. (2006). Twenty first century science: Insights from the design and implementation of a scientific literacy approach in school science. *International journal of science education*, *28*(13), 1499-1521. https://doi.org/10.1080/09500690600718344

Michal, A. L., & Shah, P. (2024). A practical significance bias in laypeople's evaluation of scientific findings. *Psychological Science, 35*(4), 315–327. https://doi.org/10.1177/09567976241231506

National Research Council. 1996. *National Science Education Standards*. Washington, DC: The National Academies Press. https://doi.org/10.17226/4962.

National Research Council. 2012. *A Framework for K-12 Science Education: Practices, Crosscutting Concepts, and Core Ideas*. Washington, DC: The National Academies Press. https://doi.org/10.17226/13165

National Research Council. 2013. *Next Generation Science Standards: For States, By States*. Washington, DC: The National Academies Press. https://doi.org/10.17226/18290

Nowak, K. H., Nehring, A., Tiemann, R., & Upmeier zu Belzen, A. (2013). Assessing students' abilities in processes of scientific inquiry in biology using a paper-and-pencil test. *Journal of Biological Education*, *47*(3), 182-188. https://doi.org/10.1080/00219266.2013.822747

OECD. (2000). *Measuring student knowledge and skills: The PISA 2000 assessment of reading, mathematical and scientific literacy*. OECD Publishing. https://doi.org/10.1787/9789264181564-en

OECD. (2016). *PISA 2015 assessment and analytical framework: Science, reading, mathematic and financial literacy*. OECD Publishing. https://doi.org/10.1787/9789264255425-en

OECD. (2023). *PISA 2025 science framework*. OECD. https://pisa-framework.oecd.org/science-2025/

Opitz, A., Heene, M., & Fischer, F. (2017). Measuring scientific reasoning—A review of test instruments. *Educational Research and Evaluation, 23*(3-4), 78–101. https://doi.org/10.1080/13803611.2017.1338586

Inhelder, B., & Piaget, J. (1958). *The growth of logical thinking: From childhood to adolescence.* (A. Parsons & S. Milgram, Trans.). Basic Books. https://doi.org/10.1037/10034-000

Rönnebeck, S., Bernholt, S., & Ropohl, M. (2016). Searching for a common ground – A literature review of empirical research on scientific inquiry activities. *Studies in Science Education, 52*(2), 161–197. https://doi.org/10.1080/03057267.2016.1206351

Shah, P., Michal, A., Ibrahim, A., Rhodes, R., & Rodriguez, F. (2017). What makes everyday scientific reasoning so challenging?. In *Psychology of learning and motivation* (Vol. 66, pp. 251-299). Academic Press. https://doi.org/10.1016/BS.PLM.2016.11.006

Tabak, I., & Dubovi, I. (2023). What drives the public's use of data? The mediating role of trust in science and data literacy in functional scientific reasoning concerning COVID-19. *Science Education, 107*(5), 1071–1100. https://doi.org/10.1002/sce.21789

Tobin, K. G., & Capie, W. (1981). The development and validation of a group test of logical thinking. *Educational and Psychological Measurement, 41*(2), 413–423. https://doi.org/10.1177/001316448104100220

Zimmerman, C. (2000). The development of scientific reasoning skills. *Developmental Review, 20*(1), 99–149. https://doi.org/10.1006/drev.1999.0497

Chapter 2 – Validity and validation in psychological measurement

Validity is the cornerstone of psychological measurement, determining how test scores can be meaningfully interpreted and used. Over time, the concept has evolved from separate types of validity toward a unified framework that views validation as an ongoing scientific argument built from multiple sources of evidence. This chapter traces that evolution, outlines the contemporary view of validity defined by the *Standards for Educational and Psychological Testing* (AERA, APA, & NCME, 2014), and concludes by translating these principles into practice through the five sources of validity evidence.

1. The Evolution of Validity Theory

Validity theory has progressively evolved from fragmented notions to a unified, evidence-based framework. The following sections trace this development from its historical foundations to contemporary perspectives.

1.1 Historical Background

The concept of validity has been central to psychological measurement since the mid-twentieth century, when psychologists began to move beyond purely predictive or content-based approaches to testing. In their seminal paper, Cronbach and Meehl (1955) introduced the notion of construct validity, arguing that tests are not validated in isolation but only in relation to the *nomological network* of theoretical constructs and observable indicators in which they are embedded. According to their formulation, a construct is admissible only if it forms part of a web of theoretical propositions linking constructs to each other and to empirical observations. Validity, therefore, was not to be understood as a single coefficient or a property inherent to a test, but as a scientific argument built through the accumulation of theoretical and empirical evidence.

Building on this, Loevinger (1957) emphasized the role of objective tests as *instruments of psychological theory*, and outlined what would later become a standard tripartite structure of validation: substantive evidence, concerning the theoretical definition and content representation of the construct; structural evidence, involving the internal consistency and dimensionality of the scale; and external evidence, situating the construct within a larger network of relationships with other constructs and relevant criteria. This perspective marked a decisive shift: tests were no longer seen as mere tools for classification or prediction, but as vehicles through which psychological theory itself could be articulated, tested, and refined.

Later developments reinforced this line of thought. Samuel Messick (1989, 1995) advanced the conceptual foundations of validity by reframing it as a property not of the test itself, but of the *inferences* made from test scores. In his view, what must be validated is not the instrument per se, but the interpretations and actions based on its scores – that is, the degree to which empirical evidence and theoretical rationales support the meaning and intended uses of those scores. This shift marked a decisive move from a test-centered to an inference-centered conception of validation, emphasizing the interpretive argument that links observed

performances to underlying constructs. As Messick (1995) clearly stated, "validity is an integrated evaluative judgment of the degree to which empirical evidence and theoretical rationales support the adequacy and appropriateness of inferences and actions based on test scores" (p. 741). The implication is that no single study, coefficient, or method can establish validity once and for all. Rather, validation is an ongoing, cumulative process – one that must be revisited whenever a test is adapted, applied to a new population, or used for a novel purpose. Building on this premise, Messick articulated a unified theory of validity that integrates all traditional "types" of validity (content, criterion, and construct) into a single framework. Within this framework, diverse sources of evidence converge to justify both the *meaning* and the *consequences* of score interpretation and use. He also underscored the inherently evaluative nature of validity, which entails not only scientific adequacy but also ethical and social responsibility in testing.

Building on Messick's inferential and interpretive framework, Zumbo and Chan (2014) made the distinction between validity and validation explicit, separating *validity*, the quality of the inferences, claims, and decisions that can be drawn from test scores; from *validation*, the process of gathering and evaluating evidence to support their appropriateness, meaningfulness, and usefulness. This distinction operationalizes Messick's argument that tests are not validated per se, but that validity resides in the interpretive claims made from their scores. This distinction is not merely semantic but has important implications for research practice: while validity refers to the interpretive soundness of test-score uses, validation captures the active process through which that soundness is established. Hence, validity represents the evaluative judgment – the *destination* – whereas validation represents the ongoing scientific process – the *journey* – that substantiates and refines that judgment. Consequently, to speak of a test as "valid" per se is misleading, because validity is always specific to a particular interpretation and context of use; what researchers engage in, instead, is *validation* – a cumulative, context-sensitive process of assembling theoretical rationale and empirical findings across settings and populations.

This historical trajectory – from Cronbach and Meehl's (1955) theoretical network, through Loevinger's (1957) tripartite framework, to Messick's unified view and its subsequent refinements by Zumbo and Chan (2014) – laid the foundation for contemporary integrative approaches to validation, where multiple strands of evidence collectively sustain the interpretive argument for test score meaning.

1.2 Contemporary View of Validity

This cumulative evolution, from early theoretical formulations to integrative and inference-centered models, has progressively shaped the contemporary understanding of validity. Building on this lineage, current frameworks translate these conceptual advances into explicit standards and applied practices for validation, emphasizing that score meaning must always be justified through converging sources of evidence and contextualized interpretations.

Building on Messick's (1989, 1995) unified theory, the contemporary view – crystallized in the *Standards for Educational and Psychological Testing* (AERA, APA, & NCME, 2014) – regards validity as a dynamic process

of justification rather than a static property of tests. It emphasizes that multiple, complementary sources of evidence must be integrated into a coherent argument about score meaning. These sources (discussed in detail in third section of this chapter) span both theoretical and empirical domains, encompassing the content, structure, and consequences of testing.

One implication is that every new use of a measure requires validation. Consider the case of adapting a scale validated in one cultural or linguistic context for use in another. This process does not merely involve linguistic translation; it requires a systematic sequence of steps: forward and back-translation, reconciliation by expert panels, and cognitive interviews to ensure semantic and cultural equivalence of items (Willis, 2004; ITC Guidelines, 2017). Once adapted, the instrument must undergo structural validation – testing dimensionality, reliability, and measurement invariance across groups (Millsap, 2011) – to ensure that scores have the same meaning across populations. Finally, external evidence is needed: convergent and discriminant validity with related constructs, predictive validity with relevant outcomes, and integration into a broader nomological network (Cronbach & Meehl, 1955). Each step exemplifies how validation is cumulative, context-bound, and never complete.

This vision is echoed in the most recent EFPA *Test Review Model* (EFPA, 2025), which incorporates the unified theory of validity while also highlighting new priorities: transparency of reporting, equity, diversity and inclusion (EDI), and the adaptation of instruments to digital formats. The model requires that reviewers explicitly document the sources of validity evidence and evaluate whether they justify the specific use of a test. This underscores the central principle that validity is not universal but situated: what holds in one population, purpose, or mode of administration may not automatically hold in another.

In doing so, the EFPA framework extends the validity argument into the ethical and technological domains, emphasizing that psychometric soundness must align with transparency, accessibility, and equity. By requiring rigorous validation for any adaptation or modification, it discourages the uncontrolled proliferation of constructs and measures, fostering cumulative progress through fewer, better validated, and more widely shared instruments.

In this evolving landscape, the shift from viewing validity as a property of tests to viewing it as a property of inferences carries profound implications for research priorities. If validity concerns the soundness of the inferences drawn from scores, then attention must also turn to *how* those scores come to be – namely, to the cognitive, contextual, and social processes underlying test responses. Understanding these response processes becomes essential to ensure that inferences remain appropriate across populations, languages, and modes of administration. In this sense, validity is inherently *situated*: it depends not only on the statistical adequacy of a measure but also on the interpretive and ethical soundness of the reasoning that connects responses to constructs.

2. Why Validated Measures Matter

The cumulative and context-sensitive nature of validation also highlights why robust measurement is essential for credible psychological research. Test validity rests not on numerical coefficients but on a coherent body of theoretical and empirical justification. When such foundations are lacking, inferences from test scores become fragile: group comparisons may be biased, constructs distorted, and theories built on unstable ground.

Recognizing this distinction also helps illuminate some of the recurrent shortcomings in psychological research. Many studies treat validity as if it were an inherited property of a scale, relying almost exclusively on internal consistency estimates such as Cronbach's alpha. In practice, this confusion between reliability and validity has contributed to a superficial evaluation of measures, where evidence about score meaning and theoretical alignment is often missing.

Some empirical reviews confirm these concerns. For example, Flake, Pek, and Hehman (2017) examined articles published in *Journal of Personality and Social Psychology* and found that nearly half of the measures had no citation to prior validation, and about 30% were single-item scales—measures inherently incapable of capturing the breadth of complex constructs. Reliability coefficients, usually alpha, were often the only psychometric information reported. Such practices not only weaken the credibility of individual studies but also contribute to the broader replication crisis in psychology, where results fail to reproduce because the measurement foundations are shaky. More recent investigations have confirmed and extended these observations. For instance, Hussey and Hughes (2020) demonstrated that many widely used scales in social and personality psychology exhibit what they termed "hidden invalidity" – structural flaws that compromise score interpretation despite apparently acceptable reliability indices. Similarly, Flake and Fried (2020) reviewed questionable measurement practices across subfields and argued that insufficient validation is among the root causes of psychology's replication and generalizability crises. As Borsboom (2006) observed, these issues reflect a deeper conceptual lag: psychology has often trailed behind psychometrics, relying on outdated practices – such as principal component analysis or Cronbach's alpha – as if statistical adequacy alone could ensure sound measurement. This overreliance on routine indices has obscured the interpretive and evidential nature of validation itself. Together, these studies converge on the same conclusion: without systematic validation, measurement remains the weakest link in the chain of psychological inference. The practical consequences of this weakness become evident when scientific claims rely on instruments of unknown or questionable validity. The issue extends well beyond methodological rigor: measurement choices shape the cumulative progress of science. When instruments are insufficiently validated or modified without documenting their psychometric implications, findings lose comparability across studies and contexts.

At the disciplinary level, these weaknesses manifest as measure proliferation and fragmentation. Large-scale analyses by Anvari and colleagues (2025) show that psychology faces an overproduction of constructs and instruments: thousands of new scales are introduced each year, most used only once, creating a fragmented

and incoherent landscape. Elson et al. (2023) vividly called this the “toothbrush problem”: every researcher prefers their own instrument, leading to an uncontrollable proliferation of measures that are rarely reused.

However, not all instrument modification contributes to this proliferation. When a measure is adapted across languages, cultures, or populations, such adaptation does not constitute the creation of a new instrument but rather an extension of its validity argument to a new context of use. In these cases, researchers must strike a balance between preserving the theoretical and structural integrity of the original scale and ensuring its semantic and cultural appropriateness in the target context. This process reflects a trade-off between rigid replication and context-sensitive adaptation: the goal is not to reinvent the measure, but to document, through systematic validation, that it continues to measure the same construct under the new conditions. Properly conducted adaptations thus contribute to cumulative progress, whereas unsystematic modifications or unvalidated derivatives exacerbate fragmentation.

Sound validation practices are therefore crucial not only for the internal integrity of individual studies but for the cumulative development of psychological science. As Loken and Gelman (2017) observed, the replication crisis is deeply intertwined with problems of measurement: when instruments yield noisy or poorly defined scores, estimates become unstable, true effects are obscured, and findings fail to replicate. Weak measurement, in their view, does not merely add random error, it undermines the very process by which scientific knowledge accumulates. Without rigorous validation, evidence cannot be meaningfully aggregated, meta-analyses lose credibility, and theoretical progress stalls. Conversely, sustained investment in validation – whether for new instruments or adapted versions of existing ones – creates the foundation for knowledge that accumulates, replicates, and ultimately informs real-world applications in education, health, and policy.

3. From Theory to Practice: The Five Sources of Validity Evidence

If validity is understood as a unified but multifaceted argument about the meaning of test scores, then the task of validation is to collect evidence from multiple, complementary sources. The *Standards for Educational and Psychological Testing* (AERA, APA, & NCME, 2014) organize these sources into five broad categories: test content, response processes, internal structure, relations to other variables, and consequences of testing. Each provides a distinct yet interconnected perspective on whether the interpretations and uses of scores are warranted. Their relevance depends on the nature and purpose of the test, but together they form the empirical foundation of the validity argument.

The following sections briefly outline these five sources as a conceptual guide rather than an exhaustive treatment. Readers interested in a comprehensive discussion are referred to the *Standards for Educational and Psychological Testing* which provides a detailed framework for designing, gathering, and interpreting validity evidence. While the relevance of each source depends on the purpose and nature of the test, together they form the empirical foundation of the validity argument.

3.1 Evidence Based on Test Content

The first source of validity evidence concerns the representativeness and relevance of test content relative to the theoretical construct. This evidence addresses whether the items, tasks, or stimuli adequately sample the conceptual domain the test purports to measure. In practice, content-based validation involves mapping the construct's components, consulting theoretical and empirical literature, and engaging subject-matter experts to evaluate content coverage and relevance. Quantitative indices and qualitative judgments are often combined to document how well the test items reflect the intended facets of the construct while avoiding irrelevant or redundant material.

For guidance on collecting and evaluating content-based evidence, readers are encouraged to consult the methodological sources that detail best practices in construct definition and content validation (e.g., Haynes et al., 1995; Lynn, 1986; Sireci, 1998). Practical applications of expert-judgment techniques and quantitative indices can be found in Lawshe (1975) for content validity ratios, and in Polit and Beck (2006), which discuss adaptations and refinements of content validity indices for psychological and educational instruments..

In the present dissertation, evidence based on test content was primarily established through an extensive review of the literature on scientific reasoning and related constructs, the results of which are presented partly in Chapter 1 (the conceptual definition and theoretical framework) and partly in Chapter 3 (the umbrella review of empirical studies). Together, these analyses delineated the domain of scientific reasoning and distinguished it from adjacent constructs.

3.2 Evidence Based on Response Processes

The second source concerns the cognitive, affective, and linguistic processes by which respondents interact with test items. Historically, response processes were viewed as a subsidiary "quality control" step, but contemporary perspectives regard them as substantive validity evidence (Zumbo & Hubley, 2017). They reveal whether items truly elicit the intended reasoning, strategies, or judgments, thus clarifying *why* items cohere beyond what factor analyses can show.

The *Standards for Educational and Psychological Testing* explicitly include response processes among the five sources of validity evidence, emphasizing their role in confirming that responses reflect the intended cognitive operations. Common methods include think-aloud protocols and cognitive interviewing, which make respondents' interpretations and decision rules explicit, helping to detect misalignments between intended and enacted reasoning (Willis, 2004). Researchers also use behavioral and paradata traces – such as response latencies, option revisits, or skipping patterns – to identify satisficing or unintended strategies (Zumbo & Hubley, 2017).

Two features make response-process evidence uniquely informative. First, it is diagnostic: it locates where score meaning breaks down (lexical ambiguity, cultural pragmatics, misleading examples, unintended cueing), guiding targeted revisions that structural analyses alone cannot specify. Second, it is context-sensitive: because

processes are shaped by language, culture, mode of administration, and stakes, they are especially critical when adapting instruments across populations or delivery formats. Indeed, the Standards make explicit that adaptation requires not only translated wording but also evidence that the same cognitive operations are being recruited in the target context.

In sum, response-process evidence does not replace internal or external validation—it grounds them, ensuring that the constructs measured align with the cognitive operations they are meant to represent.

Readers interested in methodological approaches for collecting and interpreting response-process evidence are encouraged to consult foundational works such as Ericsson and Simon (1993) and Willis (2004) on cognitive interviewing and verbal protocol analysis. For psychometric perspectives on how response processes inform construct validity, Zumbo and Hubley (2017), Padilla and Benítez (2014), and Leighton (2019) provide comprehensive treatments that integrate qualitative and quantitative approaches.

For this dissertation's aims, response-process work plays three roles. Conceptually, it tests the substantive claim that items on scientific reasoning scales genuinely elicit coordination of theory and evidence rather than superficial cues. Psychometrically, it anticipates and helps explain structural anomalies by showing where respondents collapse facets or use unintended shortcuts. Practically, in the Italian adaptation, cognitive interviewing and small-scale think-alouds will inform wording refinements, cultural substitutions for examples, and instructions that preempt common misinterpretations, thereby strengthening the bridge from item content to intended reasoning processes.

3.3 Evidence Based on Internal Structure

The third source of validity evidence concerns the internal organization of a measure, whether the relationships among items (correlations, factor loadings, and residuals) are consistent with the theoretical structure of the construct. This domain, known as structural validity, includes evidence on dimensionality, internal consistency, and measurement invariance (Loevinger, 1957; AERA, APA, & NCME, 2014).

The specific analytic approach depends on the measurement model adopted. Within classical test theory (CTT), internal structure is typically examined through exploratory and confirmatory factor analysis (EFA, CFA), which test whether items cluster as expected and whether cross-loadings or residuals suggest conceptual overlap or method effects. In item response theory (IRT), the same principles are extended to the item level: models evaluate how item parameters – difficulty, discrimination, and guessing – relate to the latent trait, providing a detailed view of how measurement precision varies across the continuum of ability or reasoning proficiency.

A closely related property is reliability, which refers to the precision or stability of measurement. It indicates the extent to which observed scores are free from random error and thus reflect consistent ordering among individuals across items or occasions (Loevinger, 1957). Reliable scores are a prerequisite for validity, because unstable scores cannot meaningfully represent any construct. Yet, as Cronbach and Meehl (1955) emphasized,

reliability alone does not guarantee that a test measures what it intends to measure. A scale can be perfectly consistent while systematically assessing the wrong construct; hence, reliability is necessary but not sufficient for validity (Messick, 1995). The most common index, Cronbach's α, assumes equal item loadings (tau-equivalence) and often misrepresents true reliability, particularly for multidimensional scales (Raykov, 1997; Sorgente & Zumbo, 2025). Modern psychometrics offers better tools. Coefficient ω (McDonald, 1999) provides a more accurate estimate of internal consistency under congeneric models, and is now widely recommended as a default. Test–retest reliability, though often neglected, remains crucial for constructs expected to be stable over time (Kline, 2023). In IRT, information functions quantify score precision across the trait continuum, revealing whether a scale is informative at low, medium, or high levels of the construct (Embretson & Reise, 2000).

Finally, measurement invariance addresses whether scores have the same meaning across groups. In classical test theory (through confirmatory factor analysis), invariance is evaluated through increasingly restrictive models: configural, weak, strong, and strict. Configural invariance tests whether the same factor structure holds across groups, that is, whether respondents conceptualize the construct similarly, with the same pattern of item–factor relations. Weak invariance constrains factor loadings to equality, ensuring that items contribute to the latent construct with the same weight, a prerequisite for comparing relationships between constructs across groups. Strong invariance further constrains intercepts (or thresholds for categorical indicators), enabling meaningful comparisons of latent means. Strict invariance additionally constrains residual variances, allowing comparisons of observed means across groups. Beyond these steps, some authors distinguish tests of structural invariance – factor variances, covariances, and means – once measurement invariance has been established (Widaman et al., 2010). Item response theory offers an analogous lens through differential item functioning analyses, which identify items that favor one group over another despite equivalent levels of the latent trait (Holland & Wainer, 1993). The practical consequences of ignoring invariance are substantial. Large-scale studies of intelligence and educational achievement have often reported group differences without first testing for invariance, raising the possibility that observed disparities reflect item bias rather than true ability differences (Byrne et al., 1989; Meredith, 1993; Millsap & Everson, 1993). Such methodological shortcuts, as Borsboom (2006) argued, carry profound ethical implications, since biased measures threaten the fairness and interpretability of test scores. Similarly, when instruments are adapted into new languages or cultures, assuming equivalence without empirical testing is untenable: wording, examples, and cultural references can alter item difficulty or interpretation, producing construct-irrelevant variance that invalidates comparisons. In sum, measurement invariance is not a technical embellishment but a validity imperative. It safeguards fairness, supports comparability across studies and populations, and anchors the claim that scores reflect the same construct regardless of who is responding. Without it, the promise of cumulative and generalizable science collapses into local and potentially biased interpretations. Internal-structure evidence therefore establishes that a measure is coherent, precise, and fair: it anchors the claim that test scores reflect the intended construct, and that this meaning holds consistently across items, individuals, and groups.

For methodological discussions of internal-structure evidence and its role in the validity argument, see the classic treatments by Messick (1989, 1995). For detailed guidance on factor analytic methods, Brown (2015) provides an applied overview of CFA, while Embretson and Reise (2000) and Reise and Revicki (2015) discuss IRT approaches to dimensionality and reliability. Millsap (2011), Putnick and Bornstein (2016), and van de Schoot et al. (2012) offer comprehensive discussions of measurement invariance, and Holland and Wainer (1993) remains the key reference for differential item functioning. Recent contributions by Dunn et al. (2014), McNeish (2018), and Kline (2023) critically review modern approaches to estimating reliability, including the use of McDonald's ω and other model-based indices.

In the empirical part of this dissertation, evidence concerning the internal structure of the Italian Scientific Reasoning Scale (SRS) is examined through a sequential and multi-method approach. Chapter 4 presents the confirmatory factor analyses conducted to test the dimensional configuration of the scale and evaluate the fit of the hypothesized measurement model, along with measurement invariance analyses assessing whether this structure holds equivalently across gender, age, and other relevant groups. Reliability is estimated through composite reliability coefficients (ω), providing further evidence of the internal consistency and homogeneity of the latent construct. Chapter 5 deepens this investigation by comparing alternative analytical frameworks – Classical Test Theory, Item Response Theory, and Latent Class Analysis. While CTT and IRT represent complementary measurement models focused on the continuous latent structure of the scale, LCA is employed in an exploratory capacity to identify potential heterogeneity in response patterns that may not be captured by traditional dimensional models. Together, these studies offer cumulative evidence on the dimensional coherence, reliability, and generalizability of SRS scores, thereby substantiating the structural phase of the overall validity argument.

3.4 Evidence Based on Relations to Other Variables

The fourth source concerns the network of theoretical and empirical relations connecting test scores to other variables. Evidence in this domain supports the interpretation of what a test measures by situating it within a nomological network (Cronbach & Meehl, 1955). Such evidence typically includes associations with theoretically related constructs, weak or null relations with distinct constructs, and systematic links with relevant behavioral or attitudinal outcomes. Together, these relations help establish whether the test behaves as theoretically expected within a broader system of constructs. These analyses are often conducted within the framework of structural equation modeling (SEM), which allows for the simultaneous estimation of latent factors and their interrelations while controlling for measurement error.

For methodological discussions on gathering and interpreting evidence based on relations to other variables, readers may consult Kane's (2013) validity argument framework, and key psychometric sources such as Borsboom and colleagues (2004) and Hubley and Zumbo (2011). For practical guidance on modeling such relations, see Little and others (2007), and Marsh and others (2014), which illustrate the use of SEM for testing construct relations while accounting for measurement error.

In the present dissertation, these principles guide the external validation of the Scientific Reasoning Scale (SRS). Using SEM, evidence supporting validity claims was gathered by examining the associations between the latent SRS factor and conceptually related constructs such as epistemic cognition, scientific literacy, and reasoning ability. These analyses test whether scientific reasoning occupies the theoretically predicted position within this conceptual network. Additional models examined associations with external outcomes to verify whether SRS scores predict theoretically relevant belief patterns beyond general cognitive or educational factors. Together, these findings provide evidence that the SRS captures a distinct and theoretically coherent dimension of reasoning.

3.5 Evidence Based on the Consequences of Testing

A contemporary account of validity is incomplete without considering the consequences of testing. Messick (1989, 1995) argued forcefully that validity encompasses not only evidence about score meaning but also the implications of how scores are interpreted and used. In his unified framework, consequences are not peripheral but integral: if a test systematically produces adverse impact, fosters misinterpretation, or legitimizes inappropriate decisions, then the validity of inferences is compromised, regardless of how well the scale fits a factor model. This consequential dimension is increasingly salient in modern testing environments. The *Standards for Educational and Psychological Testing* explicitly include the "consequences of testing" as a source of validity evidence, urging researchers and practitioners to consider both intended and unintended effects. Similarly, the *EFPA Test Review Model* (EFPA, 2025) requires reviewers to evaluate not only technical quality but also fairness, accessibility, and transparency in reporting. Within this framework, issues of equity, diversity, and inclusion (EDI) are brought to the forefront. Instruments are expected to function equivalently across subgroups, without disadvantaging individuals on the basis of language, cultural assumptions, or digital delivery formats. Moreover, the interpretations and uses of scores must remain proportionate to the empirical evidence that supports them, ensuring that testing practices are both scientifically and ethically justified. Concrete examples illustrate why this dimension matters. Consider an educational assessment that yields lower average scores for students from minority backgrounds. Without evidence of measurement invariance, we cannot know whether these differences reflect genuine disparities in the targeted competence or construct-irrelevant bias embedded in the items. Or take a clinical screener that, when adapted hastily into another language, leads to overdiagnosis in the target population. In both cases, the consequences of use – stigmatization, misallocation of resources, erosion of trust – become part of the validity argument itself. Moreover, contemporary debates about psychological measurement highlight how instruments can shape not only decisions but also constructs themselves. Anvari et al. (2025) describe how the proliferation of ad hoc measures risks reifying constructs that are poorly defined, leading to fragmented literatures where names proliferate faster than evidence. The consequence is not only inefficiency but also conceptual distortion: measures create constructs as much as they measure them. Here, responsible use of instruments means resisting the temptation to generate new scales without clear evidence of non-redundancy, and instead investing in refining and adapting those already available. In short, consequences are not an afterthought. They are the

arena where psychometric arguments meet ethical responsibility. By embedding considerations of fairness, impact, and appropriateness within the validity framework, contemporary validation recognizes that a test's legitimacy depends as much on how it is *used* as on how well it *fits*.

For comprehensive discussions on the consequences of testing and the ethical dimensions of validity, readers are referred to Messick's foundational essays (1989, 1995). Popham (1997) and Shepard (1997) provide critical perspectives on the integration of test use and impact within validity theory. Lane and Raymond (2011) offer methodological guidance for evaluating intended and unintended consequences in applied contexts, while Zumbo and Hubley (2017) extend these principles to fairness, accessibility, and equity considerations. For broader philosophical treatments, Newton and Shaw (2014) and the EFPA Test Review Model (EFPA, 2025) highlight the alignment between psychometric soundness and social responsibility in modern assessment practices.

For the present dissertation, acknowledging consequences entails two main considerations. First, it requires explicit attention to the fairness and equity of score interpretation. Although measurement invariance analyses belong to the structural phase of validation, their implications extend to this domain: demonstrating invariance across demographic and educational groups is essential to ensure that observed differences reflect genuine variation in reasoning skills rather than measurement artifacts. Second, acknowledging consequences involves reflecting on the appropriate and responsible uses of Scientific Reasoning Scale scores. The Scientific Reasoning Scale is not intended as a diagnostic tool for individuals nor as a high-stakes selection instrument; rather, its validity supports inferences about group-level patterns of reasoning, research comparisons, and educational evaluations. Making these boundaries explicit is, in itself, an integral component of the validity argument.

4. Toward Transparent and Cumulative Validation Practices

Building on the concerns raised earlier about measurement fragmentation, recent frameworks have converged on the need for transparent, cumulative, and community-centered validation practices. The *Standards for Educational and Psychological Testing* (AERA, APA, & NCME, 2014) conceptualize validity as a collective and iterative argument that grows through evidence accumulation. The EFPA *Test Review Model* (EFPA, 2025) reinforces this view by requiring explicit documentation of test adaptations, fairness evaluations, and transparent reporting of all sources of evidence. Similarly, the *SOBER guidelines* (Standardisation Of BEhavior Research; Elson et al., 2023) urge researchers to (a) demonstrate non-redundancy with existing instruments, (b) preregister scoring rules and deviations, (c) make items and scoring instructions openly available, and (d) provide sufficient information for meta-analytic synthesis.

To operationalize these principles, this dissertation was informed by the main recommendations of the *Standards* (2014), the EFPA *Test Review Model* (EFPA, 2025), the ITC Guidelines (2017), and the SOBER framework (Elson et al., 2023), seeking to integrate their principles of transparency, documentation, and cumulative reporting throughout the validation process.

Applying this framework to the Italian adaptation of the Scientific Reasoning Scale (SRS) ensures that every step – from translation choices to invariance testing, from reliability estimates to external correlates – is transparently reported. In doing so, the study aligns with current efforts to counter fragmentation by strengthening and extending an existing instrument rather than creating a new one. Ultimately, comprehensive reporting is not bureaucratic formality but the mechanism through which psychological science can evolve from dispersion toward genuine cumulativity.

Despite the fact that most of these theoretical principles on validity have been established for decades – from Cronbach and Meehl's (1955) call for construct validation to Messick's (1989) unified view and the more recent Standards (AERA, APA, & NCME, 2014) – their practical uptake in test development and use remains limited. Many practitioners and researchers still approach validity as a procedural requirement rather than as the core of measurement interpretation. This gap between conceptual advancement and applied practice underscores the urgency of restating and disseminating these ideas, making them accessible to those who design, adapt, and interpret psychological instruments.

In sum, this chapter outlined the theoretical foundations of validity as a unified and cumulative argument grounded in multiple sources of evidence. Validation is not a technical formality but an ongoing, transparent process that links theory, measurement, and ethical responsibility. The following chapters translate these principles into practice. Chapter 3 synthesizes prior research to define the construct of scientific reasoning and its content domain. Chapter 4 presents the Italian adaptation and structural validation of the Scientific Reasoning Scale (SRS), while Chapter 5 deepens structural validation by comparing complementary measurement frameworks to probe dimensionality, item functioning, reliability, and response-pattern heterogeneity. Together, these studies build a coherent validity argument for the SRS within the Italian context.

Bibliography

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

Anvari F, Alsalti T, Oehler LA, et al. (2025). A Fragmented Field: Construct and Measure Proliferation in Psychology. *Advances in Methods and Practices in Psychological Science*, 8(3). https://doi.org/10.1177/25152459251360642

Borsboom, D. (2006). The attack of the psychometricians. *Psychometrika, 71*(3), 425–440. https://doi.org/10.1007/s11336-006-1447-6

Borsboom, D., Mellenbergh, G. J., & van Heerden, J. (2004). The Concept of Validity. *Psychological Review, 111*(4), 1061–1071. https://doi.org/10.1037/0033-295X.111.4.1061

Brown, T. A. (2015). *Confirmatory factor analysis for applied research* (2nd ed.). The Guilford Press.

Byrne, B. M., Shavelson, R. J., & Muthén, B. (1989). Testing for the equivalence of factor covariance and mean structures: The issue of partial measurement invariance. *Psychological Bulletin, 105*(3), 456–466. https://doi.org/10.1037/0033-2909.105.3.456

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin, 52*(4), 281–302. https://doi.org/10.1037/h0040957

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 (London, England : 1953)*, *105*(3), 399–412. https://doi.org/10.1111/bjop.12046

Elson, M., Hussey, I., Alsalti, T. *et al.* Psychological measures aren't toothbrushes. *Commun Psychol* **1**, 25 (2023). https://doi.org/10.1038/s44271-023-00026-9

Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists.* Mahwah, NJ: Erlbaum. https://doi.org/10.4324/9781410605269

Ericsson, K. A., & Simon, H. A. (1993). *Protocol analysis: Verbal reports as data* (Rev. ed.). Cambridge, MA: MIT Press.

European Federation of Psychologists' Associations AISBL (EFPA, 2025). Model for the Review, Description and Evaluation of Psychological and Educational Tests (Test Review Model): https://www.efpa.eu/resources

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

Flake, J. K., Pek, J., & Hehman, E. (2017). Construct validation in social and personality research: Current practice and recommendations. *Social Psychological and Personality Science, 8*(4), 370–378. https://doi.org/10.1177/1948550617693063

Haynes, S. N., Richard, D. C. S., & Kubany, E. S. (1995). Content validity in psychological assessment: A functional approach to concepts and methods. *Psychological Assessment, 7*(3), 238–247. https://doi.org/10.1037/1040-3590.7.3.238

Holland, P. W., & Wainer, H. (Eds.). (1993). *Differential item functioning.* Lawrence Erlbaum Associates, Inc.

Hubley, A. M., & Zumbo, B. D. (2011). Validity and the consequences of test interpretation and use. *Social Indicators Research, 103*(2), 219–230. https://doi.org/10.1007/s11205-011-9843-4

Hussey, I., & Hughes, S. (2020). *Hidden invalidity among 15 commonly used measures in social and personality psychology. Advances in Methods and Practices in Psychological Science, 3*(2), 166–184. https://doi.org/10.1177/2515245919882903

International Test Commission. (2017). *The ITC guidelines for translating and adapting tests (2nd ed.)*. International Journal of Testing, 18(2), 101–134. https://doi.org/10.1080/15305058.2017.1398166

Kane, M. T. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement, 50*(1), 1–73. https://doi.org/10.1111/jedm.12000

Kline, R. B. (2023). *Principles and practice of structural equation modeling* (5th ed.). Guilford Press.

Lane, S., & Raymond, M. R. (2011). *Consequences of assessment and accountability systems for students and schools.* In S. Lane, M. R. Raymond, & T. M. Haladyna (Eds.), *Handbook of test development* (2nd ed., pp. 333–356). Mahwah, NJ: Erlbaum. https://doi.org/10.4324/9780203102961

Lawshe, C. H. (1975). A quantitative approach to content validity. *Personnel Psychology, 28*(4), 563–575. https://doi.org/10.1111/j.1744-6570.1975.tb01393.x

Leighton, J. P. (2019). Using think-aloud interviews and cognitive labs in educational research. *Oxford Research Encyclopedia of Education.*

Little, T. D., Card, N. A., Bovaird, J. A., Preacher, K. J., & Crandall, C. S. (2007). Structural equation modeling of mediation and moderation with contextual factors. In T. D. Little, J. A. Bovaird, & N. A. Card (Eds.), *Modeling contextual effects in longitudinal studies* (pp. 207–230). Lawrence Erlbaum Associates.

Loevinger, J. (1957). Objective tests as instruments of psychological theory. *Psychological Reports, 3*(3), 635–694. https://doi.org/10.2466/pr0.1957.3.3.635

Loken, E., & Gelman, A. (2017). *Measurement error and the replication crisis. Science, 355*(6325), 584–585. https://doi.org/10.1126/science.aal3618

Lynn, M. R. (1986). Determination and quantification of content validity. *Nursing Research, 35*(6), 382–385. https://doi.org/10.1097/00006199-198611000-00017

Marsh, H. W., Morin, A. J., Parker, P. D., & Kaur, G. (2014). Exploratory structural equation modeling: an integration of the best features of exploratory and confirmatory factor analysis. *Annual review of clinical psychology*, *10*, 85–110. https://doi.org/10.1146/annurev-clinpsy-032813-153700

McDonald, R.P. (1999). Test Theory: A Unified Treatment (1st ed.). Psychology Press. https://doi.org/10.4324/9781410601087

McNeish, D. (2018). Thanks coefficient alpha, we'll take it from here. *Psychological Methods, 23*(3), 412–433. https://doi.org/10.1037/met0000144

Meredith, W. (1993). Measurement invariance, factor analysis and factorial invariance. *Psychometrika, 58*(4), 525–543. https://doi.org/10.1007/BF02294825

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13–103). American Council on Education & Macmillan.

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), 741–749. https://doi.org/10.1037/0003-066X.50.9.741

Millsap, R. E. (2011). *Statistical approaches to measurement invariance*. Routledge. https://doi.org/10.4324/9780203821961

Millsap, R. E., & Everson, H. T. (1993). Methodology review: Statistical approaches for assessing measurement bias. *Applied Psychological Measurement, 17*(4), 297–334. https://doi.org/10.1177/014662169301700401

Newton, P., & Shaw, S. (2014). *Validity in educational & psychological assessment*. SAGE Publications Ltd, https://doi.org/10.4135/9781446288856

Padilla, J. L., & Benítez, I. (2014). Validity evidence based on response processes. *Psicothema, 26*(1), 136–144. https://doi.org/10.7334/psicothema2013.259

Polit, D. F., & Beck, C. T. (2006). The content validity index: are you sure you know what's being reported? Critique and recommendations. *Research in nursing & health*, *29*(5), 489–497. https://doi.org/10.1002/nur.20147

Popham, W. J. (1997). Consequential validity: Right concern—wrong concept. *Educational Measurement: Issues and Practice, 16*(2), 9–13. https://doi.org/10.1111/j.1745-3992.1997.tb00586.x

Putnick, D. L., & Bornstein, M. H. (2016). Measurement invariance conventions and reporting: The state of the art and future directions for psychological research. *Developmental Review, 41*, 71–90. https://doi.org/10.1016/j.dr.2016.06.004

Raykov, T. (1997). Scale reliability, Cronbach's coefficient alpha, and violations of essential tau-equivalence with fixed congeneric components. *Multivariate Behavioral Research, 32*(4), 329–353. https://doi.org/10.1207/s15327906mbr3204_2

Reise, S. P., & Revicki, D. A. (Eds.). (2015). *Handbook of item response theory modeling: Applications to typical performance assessment.* Routledge/Taylor & Francis Group.

Shepard, L. A. (1997). The centrality of test use and consequences for validity. *Educational Measurement: Issues and Practice, 16*(2), 5–8. https://doi.org/10.1111/j.1745-3992.1997.tb00585.x

Sireci, S.G. The Construct of Content Validity. *Social Indicators Research 45*, 83–117 (1998). https://doi.org/10.1023/A:1006985528729

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

van de Schoot, R., Lugtig, P., & Hox, J. (2012). A checklist for testing measurement invariance. *European Journal of Developmental Psychology, 9*(4), 486–492. https://doi.org/10.1080/17405629.2012.686740

Widaman, K. F., Ferrer, E., & Conger, R. D. (2010). Factorial invariance within longitudinal structural equation models: Measuring the same construct across time. *Child Development Perspectives, 4*(1), 10–18. https://doi.org/10.1111/j.1750-8606.2009.00110.x

Willis, G. B. (2004). *Cognitive interviewing: A tool for improving questionnaire design.* Thousand Oaks, CA: Sage. https://doi.org/10.4135/9781412963947.n73

Zumbo, B. D., & Chan, E. K. H. (Eds.). (2014). *Validity and validation in social, behavioral, and health sciences.* Springer International Publishing/Springer Nature. https://doi.org/10.1007/978-3-319-07794-9

Zumbo, B. D., & Hubley, A. M. (Eds.). (2017). *Understanding and investigating response processes in validation research.* Springer International Publishing/Springer Nature. https://doi.org/10.1007/978-3-319-56129-5

Chapter 3 - Scientific Reasoning: An Umbrella Review Across Substantive, Structural, and External Validation Phases[1]

Abstract: Scientific reasoning is a critical educational goal, essential for fostering inquiry-based science learning and preparing students for informed participation in society. Despite its centrality in global education frameworks, the concept has been defined and assessed in diverse ways, making theoretical development and empirical evaluation challenging. This umbrella review synthesizes the findings of systematic reviews and meta-analyses in order to examine the conceptualization, measurement and validation of scientific reasoning. In addition, it considers related constructs, including scientific inquiry, scientific literacy and science process skills, in order to clarify their interrelations. Following PRISMA guidelines, a systematic search identified nine eligible reviews published between 2016 and 2025. Data extraction and thematic synthesis were organized according to the three-phase model of construct validation: substantive (conceptualization), structural (measurement tools and psychometric properties), and external (validity evidence). The findings point to considerable conceptual ambiguity, a lack of unified theoretical models, and wide variability in instruments, with limited psychometric reporting and almost no evidence of external validity. Overall, the review underscores the urgent need for clearer theoretical frameworks and more rigorous validation practices—prerequisites for advancing research, supporting effective instruction, and informing educational policy on a global scale.

Keywords: umbrella review; scientific reasoning; scientific inquiry; scientific literacy; science process skills; construct validation

[1] *This chapter was co-authored by Rossella Caliciuri, Giulia Moncalieri, and Margherita Lanz, and is currently under review in Current Psychology.*

1. Introduction

1.1 Research problem and research question

Scientific reasoning is a set of cognitive skills essential to scientific inquiry and knowledge construction; it involves activities such as identifying problems, formulating and testing hypotheses, and evaluating evidence (blinded, under review; Bao et al., 2022). Scientific reasoning skills are considered a fundamental goal in science education (OECD, 2019). These skills reflect a form of higher-order thinking that is essential for 21st-century learning, as they play a key role in preparing students to address complex challenges in a world increasingly influenced by science and technology (Fatimah et al., 2025). It enables individuals to apply the scientific method, draw logical inferences, and make informed decisions (Díaz et al., 2023).

The importance of scientific reasoning is consistently emphasized across major educational frameworks, reflecting a global commitment to promoting inquiry-based science learning (Vo & Simmie, 2025). Although these frameworks originate from diverse cultural and policy contexts, initiatives such as the 'National Science Education Standards' (NRC, 2012) in the United States, the 'Twenty First Century Science' program (Millar, 2006) in England, the European Commission's 'Science Education for Responsible Citizenship' report (European Commission & Directorate-General for Research and Innovation, 2015), and the internationally implemented 'Programme for International Student Assessment' (PISA; OECD, 2023), share several foundational principles. First, they all emphasize scientific reasoning as a key set of skills for 21st-century learners, highlighting its role not only in academic success but also in informed decision-making and problem-solving. Second, these frameworks promote inquiry-based learning approaches in which students actively engage in posing questions, designing investigations, evaluating evidence, and constructing explanations. Third, they underline the importance of connecting science to real-world contexts, encouraging learners to apply scientific knowledge to everyday situations, public discourse, and global challenges. Finally, these initiatives increasingly reflect the view that science education should serve a civic function, fostering scientific literacy as a means of empowering individuals to critically assess information and engage in societal debates.

Despite its significance, there is no universally accepted definition of scientific reasoning. Existing theoretical conceptualizations focus on different aspects; in particular, a study by Engelmann et al. (2016) identifies three main strands. The first strand views scientific reasoning as a process of scientific discovery, involving activities such as formulating questions and hypotheses, designing investigations, collecting and analyzing data, and drawing conclusions (Klahr & Dunbar, 1988; Lawson, 1995; Fischer et al., 2014). The second strand conceptualizes scientific reasoning as scientific argumentation, emphasizing claims supported by evidence and logical warrants, considered both in terms of structure and as a dialogic process leading to reasoned conclusions (Toulmin, 1958; Kuhn, 1991; Osborne, 2010). Finally, the third strand approaches scientific reasoning as an understanding of the nature of science itself, including its assumptions, values, and core constructs such as hypotheses, models, and theories (NRC, 2013). Overall, these theoretical conceptualizations highlight

scientific reasoning as a complex, iterative, and non-linear set of skills, rather than a fixed or uniform process (Engelmann et al., 2016).

This variability in theoretical conceptualizations has directly shaped how scientific reasoning is measured. Numerous measurement instruments have been developed, each based on different theoretical assumptions and populations. As a result, measurement instruments vary widely in structure and purpose, limiting their generalizability and comparability across contexts. For instance, a 2017 review by Opitz et al. examined 38 different instruments intended to assess scientific reasoning, revealing a lack of clarity regarding the range of available tests, the specific skills being measured, and their psychometric properties. Furthermore, it is often unclear whether these instruments have undergone a proper process of construct validation – that is, the systematic gathering of validity evidence to support the interpretation of scores as meaningful representations of a psychological construct (Cronbach & Meehl, 1955; Flake et al., 2017). As Flake et al. (2017) emphasize, 'valid measurement is a necessary prerequisite to the interpretation of results' (p. 375). In the absence of rigorous validation procedures, there is a risk of ambiguity regarding the extent to which a given instrument can be relied upon to measure its intended parameters. This compromises not only the reliability of the findings, but also their replicability and theoretical contribution.

According to the Standards for Educational and Psychological Testing (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education (AERA, APA, & NCME, 2014), construct validation is typically approached through three interconnected phases: (1) the substantive phase, which involves developing a strong theoretical foundation and clearly defining the construct; (2) the structural phase, which examines both the types of instruments used to measure the construct (e.g., scales, tasks, tests) and evaluates their psychometric properties (e.g., factor structure, internal consistency); (3) and the external phase, which explores how the construct relates to other variables or predicts relevant outcomes (Cronbach & Meehl, 1955; Loevinger, 1957; Flake et al., 2017). By delineating the process of construct validation, these phases offer a valuable framework for examining how scientific reasoning has been defined, how it has been measured, and what forms of validity evidence have been provided.

Alongside the growth of theoretical conceptualizations and measurement instruments, there has also been an increase in reviews focusing on scientific reasoning. To get an initial sense of this trend, we conducted a preliminary search in three databases (Scopus, APA PsycInfo, and Web of Science) using keywords related to review types and scientific reasoning. This search returned 102 records, which were reduced to 63 unique entries after removing duplicates (see Figure 1).

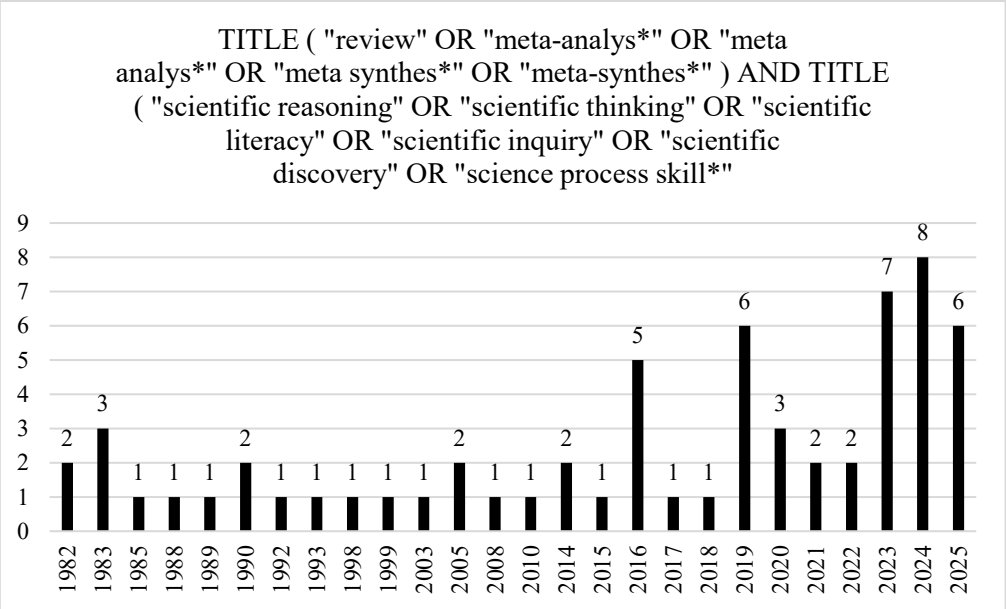

Fig. 1 *Number of Review Articles on Scientific Reasoning and Related Constructs Published per Year (1982–2025)*

Therefore, given the growing interest in scientific reasoning across different educational frameworks worldwide, the lack of a shared theoretical conceptualizations, and the wide variety of measurement instruments used to measure it, it is evident that a more comprehensive and lucid synopsis of the domain is imperative. Although numerous reviews have been published on the topic, no study to date has systematically synthesized them to examine whether (1) scientific reasoning is explicitly conceptualized; (2) which types of instruments are used to assess it and whether these instruments have been evaluated in terms of their psychometric properties (e.g., factor structure, internal consistency); and (3) whether broader validation evidence has been reported, including their relationship to other constructs or outcomes. In order to address this lacuna, an umbrella review was undertaken with the aim of mapping the extant body of literature and offering a comprehensive synthesis of key aspects within this field. The rationale for adopting this approach is rooted in the substantial proliferation of systematic and scoping reviews that have examined related issues. While this expanding collection of secondary evidence has yielded valuable insights, it remains fragmented, lacking an overarching synthesis that systematically integrates findings on definitions, measurement instruments, and validation practices. By consolidating and critically appraising the available reviews, the present umbrella review seeks to provide a higher-order perspective that will advance conceptual clarity and methodological rigor in this area.

This umbrella review is guided by the following primary research questions, which informed the design of the search strategy, inclusion/exclusion criteria, data extraction, and synthesis process. These questions are organized according to the three phases of construct validation as outlined in the Standards for Educational and Psychological Testing (AERA, APA, & NCME, 2014):

(1) Substantive phase – Conceptualization of scientific reasoning: *How is scientific reasoning conceptualized across systematic reviews and meta-analyses?*

(2) Structural phase – Measurement instruments and psychometric evaluation: *What types of instruments (e.g., scales, tasks, tests) are used to assess scientific reasoning? What are their structural characteristics (e.g., number of items, response formats, skills assessed), and what psychometric properties are reported (e.g., factor structure, internal consistency)?*

(3) External phase – Evidence of validity: *What forms of external validity evidence are reported for these instruments? Specifically, how does scientific reasoning relate to other constructs?*

Together, these questions aim to provide a critical and integrated overview of how scientific reasoning is currently conceptualized, measured and validated, highlighting strengths and limitations in the literature and informing future research, educational practice, and policy development.

2. Methods

2.1 Protocol registration

This umbrella review was conducted in accordance with the PRISMA guidelines for systematic reviews (Page et al., 2021). The review protocol was preregistered on the Open Science Framework (OSF) to ensure transparency, traceability, and reproducibility. The protocol, publicly accessible at https://osf.io/jvye7, outlines in detail the objectives, inclusion and exclusion criteria, search strategy, study selection process, data extraction methods, and planned synthesis approaches.

2.2 Search Strategy

We conducted a comprehensive search across multiple databases, including Scopus, PsycINFO, and Web of Science. To identify articles relevant to our area of interest, we used search strings consisting of all possible combinations of the terms 'scientific reasoning', 'scientific thinking', 'scientific literacy', 'scientific inquiry', 'scientific discovery', or 'science process skill'* within the title. In line with previous reviews on this topic (e.g., Edelsbrunner & Dablander, 2019; Opitz et al., 2017), we employed this variety of terms because the skills under investigation are encompassed by concepts that are referred to by different names. Additionally, to locate systematic reviews and/or meta-analyses, we included the following terms in the title: 'review', 'meta-analys'*, 'meta analys'*, 'meta synthes'*, and 'meta-synthes'*. No filters were applied regarding language, publication date, or disciplinary area, to ensure the broadest possible coverage of the existing literature. We acknowledge that searching in title, abstract, and keywords (TAK) is a common recommendation, and we began our strategy development with this scope. However, in preliminary testing we found that extending the search to abstracts and keywords yielded a very large number of non-systematic papers in which review of the literature was part of the introduction or background but not the study design. Such records substantially increased screening noise without identifying additional systematic reviews or meta-analyses that met our inclusion criteria.

Restricting the search to the title field therefore maximized specificity and feasibility, while still capturing the relevant systematic reviews and meta-analyses in this domain.

2.3 Study Screening and Selection

In accordance with the methodological standards set forth in the JBI [Joanna Briggs Institute] Manual for Evidence Synthesis (Aromataris et al., 2024), which specifies that for the purposes of an umbrella review the term *studies* refers exclusively to syntheses of research evidence—such as systematic reviews and meta-analyses—this review will include only research syntheses that meet these criteria. Only reviews that synthesize empirical evidence using recognized systematic approaches will be considered. Reviews based solely on theoretical frameworks, expert opinions, or non-empirical commentary will be excluded, in line with JBI guidance. Two additional exclusion criteria were applied: (1) reviews that did not specifically focus on scientific reasoning (or synonyms), and (2) reviews that lacked a dedicated section on measurement tools or instruments. In summary, one eligibility criterion pertained to the study design (i.e., the inclusion of systematic reviews or meta-syntheses), while the other concerned the specific topic (i.e., scientific reasoning and the instruments used to assess it).

As reported in Figure 2, our search – conducted on June 4, 2025 – yielded 102 results across Scopus (n = 51), Web of Science (n = 33) and PsycInfo (n = 18). After removing duplicates (n = 39), 63 articles were identified. The complete list is available in the supplementary materials (see *File 1* in the Excel file 'Records, Eligibility, Data') on OSF.

Upon retrieving the articles, both abstract and full-text screenings resulted in the exclusion of 54 records for the following reasons: (1) 31 articles were excluded because they could not be identified as systematic reviews or did not adopt a systematic approach; (2) 9 articles were excluded because they did not specifically address scientific reasoning; (3) 14 articles were excluded because they did not review the instruments used to measure scientific reasoning. After excluding these articles, the total number of articles included in the review was 9.

The selection process from 63 records to 9 articles was carried out by two authors (RC and GM). These authors independently read all the abstracts and evaluated each record eligibility. The agreement between independent coders was calculated through Cohen's Kappa coefficient (Cohen, 1960), where higher values (range 0-1) suggest a more robust agreement. The agreement was greater than .90, indicating an excellent level of agreement beyond chance. The complete list of articles that underwent the selection process, together with the reasons for exclusion, is provided the supplementary materials (see *File 2* in the Excel file 'Records, Eligibility, Data') on OSF. In some cases, the entire article was reviewed to ensure an accurate evaluation of the record. Any disagreements were resolved through consensus during team meetings, with the involvement of a third reviewer (ML) when necessary.

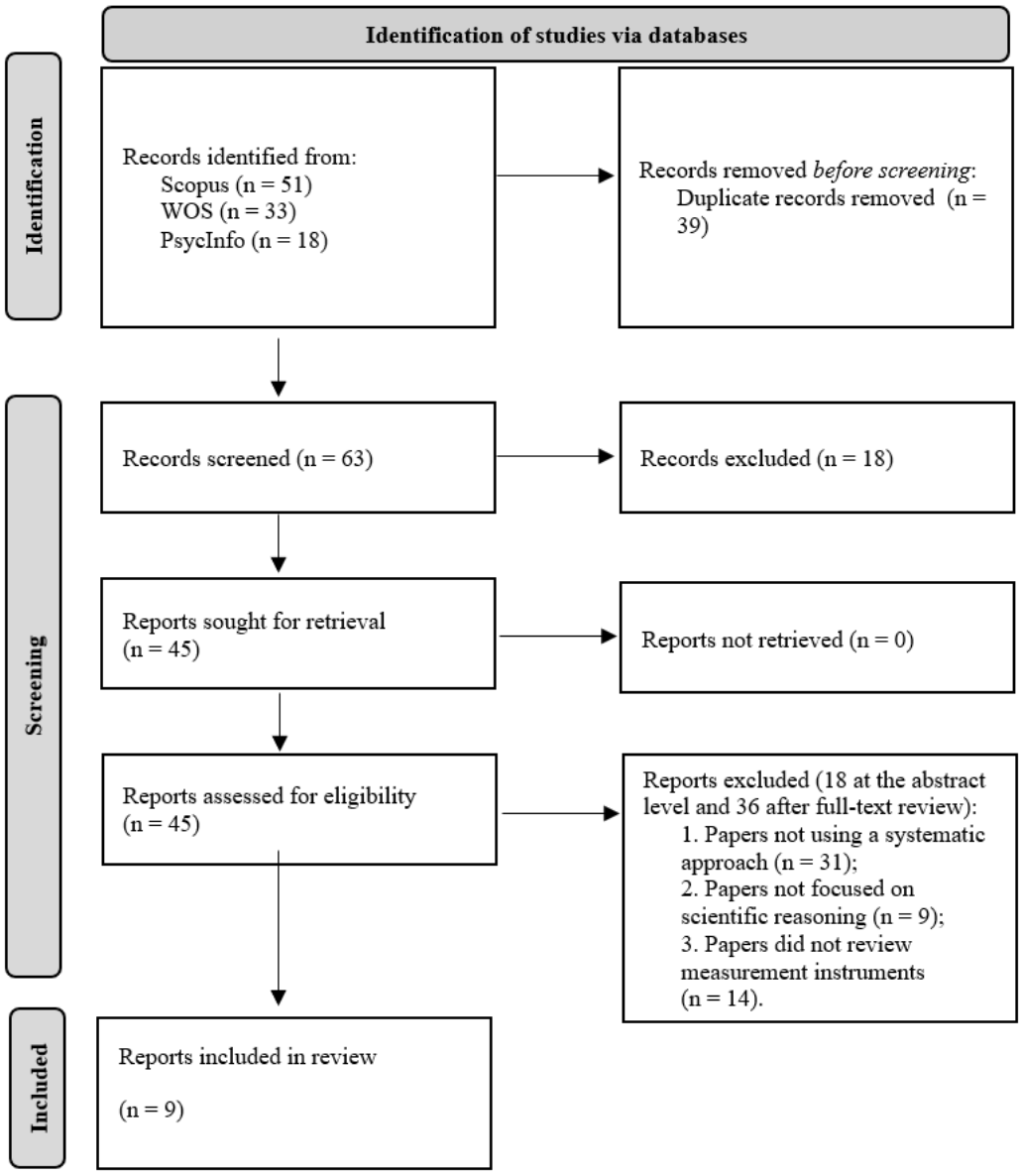

Fig. 2 *PRISMA flow diagram illustrating the flow of information at each stage of the review process*

2.4 Data Collection

Data collection refers to the process by which relevant information and data are extracted from the systematic reviews and meta-analyses included in this umbrella review. The extraction template was iteratively refined during the initial coding of the first set of articles. This piloting phase allowed the research team to discuss and test the form to ensure consistent application and to maximize the likelihood of accurately capturing information relevant to the objectives of the umbrella review. Without such preliminary piloting and team discussions, reviewers might have interpreted fields differently or assessed their relevance to the research questions inconsistently, potentially leading to unnecessary additional work to resolve discrepancies later. Importantly, due to the breadth and variability of the reviews included, sections relating to results were not coded using a rigid quantitative extraction template. Instead, they were examined in depth through comprehensive qualitative reading to capture nuances that might otherwise be overlooked by a structured form.

Once the extraction template reached its final version, all included reviews were coded by a single reviewer (RC). To ensure rigor and mitigate potential bias associated with single coding, the reviewer regularly consulted with the broader research team to discuss and resolve any uncertainties or ambiguities.

In accordance with JBI guidance, data extraction in umbrella reviews should not extend to retrieving or reanalyzing primary studies included within the systematic reviews. Thus, this umbrella review relied exclusively on the findings reported by the included syntheses, without contacting authors of primary studies or seeking additional unpublished data.

The complete results of the data extraction process are available in the supplementary materials (see *File 3* in the Excel file 'Records, Eligibility, Data') on OSF.

From each record we have extracted the following information:

1. *Publication details* (e.g., year and journal of publication, database from which the record was retrieved).
2. *Study details* (e.g., study design, number of databases consulted and searched type of review, search period in the databases, aim of the review, research questions, time frame of the included studies, number and type of primary studies, country of origin of primary studies; tool used for quality assessment of the primary studies, quality related judgment).
3. *Conceptualization of scientific reasoning* used in the study.
4. *Information on the measurement instruments* used to measure scientific reasoning (e.g., name, purpose, number of items, response format) *and psychometric evaluation* (e.g., factor structure; internal consistency).
5. *Evidence of validity* (e.g., convergent, discriminant, criterion-related).

We have also included a dedicated assessment template (presented in the next section) to evaluate the quality of the studies included in the review.

2.5 Assessment of methodological quality

The selected articles will be critically appraised for methodological quality using the standardized instrument from JBI, namely the 'JBI Critical Appraisal Checklist for Systematic Reviews and Research Syntheses' (Aromataris et al., 2024). This tool has been integrated into our template and consists of 11 items, each of which can be rated as 'Yes', 'No', 'Unclear', or 'Not Applicable'. Some example items include: 'Is the review question clearly and explicitly stated?' or 'Were the inclusion criteria appropriate for the review question?'. Any disagreements between reviewers will be resolved through discussion within the research team. The results of the critical appraisal will be reported in the supplementary materials hosted on OSF (see *File 3* in the Excel file 'Records, Eligibility, Data'). All studies, regardless of the outcome of their methodological appraisal, will

proceed to data extraction and synthesis. However, the results of the critical appraisal will inform the interpretation of findings and the strength of the conclusions drawn. Studies with significant methodological limitations will be discussed with appropriate caution, and their limitations will be explicitly acknowledged in the synthesis.

2.6 Data summary

This umbrella review synthesized data from systematic reviews and meta-analyses to provide a comprehensive mapping of how scientific reasoning is conceptualized, measured, and validated across different contexts. The approach to summarizing findings was structured to closely align with the objectives, research questions, and the data extraction process. Since the included reviews differed considerably in their aims, a narrative and thematic synthesis was employed. This approach allows for an in-depth exploration of conceptual differences, as well as a critical comparison across reviews (Gough et al., 2018).

To provide essential context for interpreting the results, the synthesis will begin with a descriptive overview of the included studies. This initial section will summarize publication details, such as the year and journal of publication and study characteristics, including the study design (e.g., systematic review, meta-analysis), the number of databases searched, the total number and types of primary studies included, and any tools used to assess the quality of those primary. Following this overview, the synthesis will be structured around three main sections, each aligned with the objectives of this umbrella review.

In the first section (*Substantive Phase – Conceptualization of Scientific Reasoning*), we will explore how scientific reasoning is conceptualized across the included reviews. In the second section (*Structural Phase – Measurement Instruments and Psychometric Evaluation*), we will examine the characteristics of the instruments used to assess scientific reasoning, as well as their psychometric properties. This part will provide a detailed account of the types of tools identified (e.g., scales, performance-based tasks, standardized tests), their key features (e.g., number of items, response formats, domains assessed, intended purposes), and reported psychometric qualities (e.g., factor structure, internal consistency). In the third section (*External Phase – Evidence of Validity*), the synthesis will address the external validity of the instruments. Specifically, it will explore the types of validity evidence reported (e.g., convergent, discriminant, criterion-related).

The studies included in this review refer to scientific reasoning using a range of associated terminology. This variation is consistent with previous reviews on the topic, where terms such as *scientific inquiry*, *science process skills*, and *scientific literacy* were explicitly included as synonyms for *scientific reasoning* in the search strings (e.g., Edelsbrunner & Dablander, 2019; Opitz et al., 2017). Following this precedent, we organized the studies based on the primary term used in each review: *scientific reasoning* (Edelsbrunner & Dablander, 2019; Opitz et al., 2017), *scientific literacy* (Istyadji & Sauqina, 2023; Roy et al., 2025), *science process skills* (Fugarasti et al., 2019; Gizaw & Sota, 2023; Yıldırım et al., 2016), and *scientific inquiry* (Rönnebeck et al., 2016; Vo & Simmie, 2025).

Overall, this layered structure of the data summary is designed to reflect the diversity of conceptualizations, measurement instruments, and validity evidence provided in the existing body of systematic reviews and meta-analyses. By organizing the synthesis in this way, we aim to provide clear, actionable insights that can inform future research directions, educational interventions, and policy considerations in the field of scientific reasoning assessment.

2.6.1 Overlap across reviews

To assess the degree of overlap among the reviews included, we first examined whether the list of primary studies was transparently reported. Out of the nine reviews, six provided a complete list of their included studies, either directly in the text (Opitz et al., 2017; Edelsbrunner & Dablander, 2019; Fugarasti et al., 2019; Istyadji & Sauqina, 2023) or in supplementary materials (Rönnebeck et al., 2016; Vo & Simmie, 2025). For Yıldırım et al. (2016), the list was not publicly available but was obtained from the authors upon request, while for two reviews (Gizaw & Sota, 2023; Roy et al., 2025) it could not be retrieved. Given this variability, we calculated the Corrected Covered Area (CCA; Pieper et al., 2014) only for the subset of seven reviews with accessible lists. The CCA is a standardized index that quantifies the degree of overlap between primary studies across reviews, adjusting for the number of included studies and reviews. By restricting the analysis to verifiable data, we ensured a transparent and conservative estimate of overlap, while acknowledging that some degree of additional overlap may exist with the two reviews lacking accessible lists.

3. Results

3.1 Publication details

The nine studies included in this umbrella review, summarized in Table 1, were published between 2016 and 2025. All were systematic reviews, except for one meta-synthesis (Yildirim et al., 2016). All articles, except for one (Fugarasti et al., 2019), were published in journals indexed by Scopus, categorized under 'education' within 'social sciences'. The number of databases searched across these reviews ranged from 1 to 14. Each study clearly stated its objectives and specific research questions. All included articles were deemed to meet an acceptable level of methodological quality based on the JBI Critical Appraisal Checklist and thus were considered qualitatively valid for inclusion in the synthesis. The checklist was applied item by item, but it does not prescribe a formal scoring system or cut-off values to categorize quality (Aromataris et al., 2024). Accordingly, we did not adopt numerical thresholds but reported the results narratively. Notably, however, only one review (Roy et al., 2025) explicitly reported using a tool to assess the quality of its included studies. For the seven reviews where the list of primary studies was available, the total number of study inclusions was 366, corresponding to 331 unique primary studies. With seven reviews considered, the Corrected Covered Area (CCA; Pieper et al., 2014) was equivalent to 1.76%. According to established thresholds, this value indicates slight overlap (0–5%) among primary studies, suggesting that the reviews synthesized largely distinct bodies of evidence. The overlap matrix used for this calculation is available in the supplementary materials on OSF (see the Excel file ' CCA -overlap').

Table 1 - *Summary of Included Reviews*[a]

Study (Author, Year)	Study Design	No. of Databases	No. of Primary Studies
Rönnebeck et al. (2016)	Systematic review	2	81
Yildirim et al. (2016)	Meta-synthesis	14	200
Opitz et al. (2017)	Systematic review	3	38
Edelsbrunner & Dablander (2019)	Systematic review	2	11
Fugarasti et al. (2019)	Systematic review	3	19
Gizaw & Sota (2023)	Systematic review	6	56
Istyadji & Sauqina (2023)	Systematic review	1	43
Roy et al. (2025)	Systematic review	3	79
Vo & Simmie (2025)	Systematic review	2	63

[a]*Note*. Full citations for each study are included in the reference list.

3.2 Conceptualization

3.2.1 Scientific Reasoning

The term scientific reasoning appears explicitly in two articles (Edelsbrunner & Dablander, 2019; Opitz et al., 2017). It is recognized as a foundational element of scientific literacy (Opitz et al., 2017). While scientific literacy broadly includes the knowledge, skills, attitudes, and civic engagement needed to navigate science in everyday life, scientific reasoning refers more specifically to the cognitive skills essential for conducting scientific inquiry and ultimately for achieving scientific literacy.

Edelsbrunner & Dablander (2019) define scientific reasoning as the reasoning and problem-solving skills involved in generating, testing, and revising hypotheses or theories. These authors emphasize that while these skills stem from everyday thinking, scientific reasoning is far more complex, highly structured, and refined. It requires systematic methodologies and rigorous evidence evaluation (Zimmerman & Klahr, 2018).

Opitz et al. (2017) identify three key conceptual debates about scientific reasoning: (a) whether it is a general, uniform competence or made up of distinct dimensions; (b) which specific skills it includes; (c) and whether it is domain-general or domain-specific.

Theories differ on whether scientific reasoning is a single, broad competence or a set of distinct dimensions (a). Early conceptualizations, beginning with Piaget & Inhelder (1958), regarded it as a unified, complex competence. In contrast, more recent models – such as Klahr & Dunbar's (1988) 'Scientific Discovery as Dual Search' – describe it as a multifaceted process composed of several distinct yet interconnected phases: hypothesis generation, evidence production, and evidence evaluation. Kuhn (1989) similarly characterized scientific reasoning as a broad process centered on coordinating theory and evidence. In line with this view, Fischer et al. (2014) conceptualize it as a multidimensional construct made up of multiple interrelated skills.

Regarding the included skills (b), Opitz et al. (2017), drawing on Fischer et al. (2014), proposed an eight-skill model, comprising: (1) problem identification: recognizing discrepancies and analyzing situations; (2) questioning: formulating inquiries to drive reasoning; (3) hypothesis generation: developing plausible answers based on models or evidence; (4) construction/redesign of artifacts: creating, testing, and refining artifacts; (5) evidence generation: gathering empirical data through experiments or observations; (6) evidence evaluation: critically analyzing evidence relative to claims; (7) drawing conclusions: weighing evidence to confirm or revise claims; (8) communicating and scrutinizing: sharing and discussing methods and results.

Finally, the debate also concerns whether scientific reasoning is domain-general or domain-specific (c). Over time, there has been a clear shift from viewing scientific reasoning as a domain-general competence toward recognizing its domain-specific aspects (Opitz et al., 2017). While Klahr and Dunbar's (1988) model did not entirely dismiss domain-general components, it highlighted the crucial role of domain-specific prior knowledge in shaping scientific reasoning. More recent models, such as Fischer et al. (2014), reject the idea of a fixed or linear order of skills in scientific inquiry. Instead, they highlight the flexible nature of scientific reasoning, where different skills can be used in varying sequences or even at the same time.

Both Edelsbrunner & Dablander (2019) and Opitz et al. (2017) agree that research differs greatly in how it defines the key components of scientific reasoning. These components are referred to in various ways (sub-competencies, primary skills, ability components or multi-skill processes). Despite this variation, there is a noticeable trend toward viewing scientific reasoning as a multidimensional construct, an integrated collection of related skills rather than a single, unified ability. Nevertheless, the debate remains open regarding whether this ability is general or specific to particular fields.

3.2.2 Scientific Literacy

Two recent reviews examine the concept of scientific literacy: Istyadji & Sauqina (2023) and Roy et al. (2025). The concept of scientific literacy has been studied since the 1950s (e.g., Hurd, 1958), driven by the growing need to understand how science and technology influence decision-making in democratic societies (Millar, 1997). Despite its long history, it still lacks a single, clear definition. Nevertheless, scientific literacy is widely recognized as a central educational goal and an essential competence for citizens in the 21st century (Laugksch, 2000). Since 2006, it has also been regularly assessed by the OECD through the PISA framework.

Over the years, various conceptualizations have been proposed (Istyadji & Sauqina, 2023; Roy et al., 2025). Miller (1983) identified three key dimensions: awareness of scientific norms and methods, cognitive knowledge of scientific content, and attitudes toward science. Shamos (1995) distinguished between cultural, functional, and true scientific literacy, the latter involving deeper theoretical and epistemological understanding. The OECD (2019), meanwhile, defines scientific literacy in terms of three core competencies: explaining scientific phenomena, evaluating and designing scientific inquiry, and interpreting scientific data and evidence. Building on this, Istyadji & Sauqina (2023) describe scientific literacy as the capacity to engage thoughtfully with science-related issues as reflective citizens, highlighting its evolution from the simple acquisition of knowledge to a broader competence needed to address complex public issues. Similarly, Roy et al. (2025) describe scientific literacy as a multidimensional construct involving conceptual understanding, the application of knowledge, inquiry skills, epistemological awareness, and attitudes toward science.

In addition to these definitional perspectives, two main visions of scientific literacy have been widely discussed in literature. First, Vision I, developed and later expanded by Roberts and Bybee (2014), emphasizes scientific content and propositional knowledge, often associated with traditional school science and focused on the products of science. In contrast, Vision II promotes the use of scientific knowledge in real-world contexts and aims to equip individuals with the ability to apply science meaningfully in everyday life. This second vision is seen as a key challenge for 21st-century science education, as it strengthens analytical skills for socio-scientific decision-making and supports a broader understanding of the social value of science (Holbrook & Rannikmae, 2009). In this perspective, scientific inquiry plays a central role in fostering scientific literacy. It encompasses both scientific practices and epistemological dimensions, with growing attention to the meaning, relevance, and real-world contexts of inquiry (Schwartz et al., 2023).

3.2.3 Science Process Skills (SPS)

Three reviews have examined the concept of science process skills (SPS), highlighting their relevance in science education and inquiry-based learning (Fugarasti et al., 2019; Gizaw & Sota, 2023; Yıldırım et al., 2016). Across these studies, SPS are consistently defined as core abilities that scientists use to investigate the natural world and that students must develop to effectively engage in scientific inquiry and construct scientific knowledge.

Although there is general agreement on the importance of SPS, definitions vary across literature. Fugarasti et al. (2019) describe them as a set of transferable skills reflecting the behaviors of scientists. Gizaw & Sota (2023) draw attention to the variety of labels used to refer to these skills, including 'procedural skills', 'experimental habits of mind', and 'scientific inquiry abilities' (Chakraborty & Gillian, 2022). Ozgelen (2012), as cited by Gizaw & Sota (2023), expands this definition further by including both mental and physical activities involved in collecting and organizing data, making predictions, explaining phenomena, and solving problems.

Despite these definitional variations, a key area of agreement across the three reviews concerns the hierarchical organization of SPS into two broad categories: Basic Process Skills and Integrated Process Skills. As outlined by Yıldırım et al. (2016) and Gizaw & Sota (2023), Basic Process Skills – typically introduced in primary education – include observing, classifying, communicating, measuring, predicting, and inferring. Integrated Process Skills, more commonly developed at the secondary level, involve more complex abilities built upon the basic ones, such as controlling variables, formulating hypotheses, interpreting data, operationally defining variables, experimenting, and modeling.

Beyond this central categorization, each review adds further interpretative layers. Gizaw & Sota (2023) revisit Finley's (1983) characterization of SPS, emphasizing their hierarchical structure, their nature as specific intellectual skills universally employed by scientists, and their applicability across content areas to support rational thinking. Fugarasti et al. (2019) stress that SPS are not isolated from content knowledge but are developed through experience and are integral to the understanding of scientific concepts. Yıldırım et al. (2016) highlight the close link between SPS and scientific applications, advocating for their explicit integration into curricula to support science learning, promote analytical thinking, facilitate knowledge construction, and foster lifelong learning.

While all three reviews emphasize the centrality of SPS, Gizaw and Sota (2023) highlight their crucial role in achieving scientific literacy. They also note that SPS are often overlooked in content-driven instruction (Coil et al., 2010) and stress the need for explicit instructional strategies to support their effective development (Abd Rauf et al., 2013).

In conclusion, the three reviews converge in describing SPS as a structured set of both basic and integrated skills that are essential not only for conducting scientific inquiry, but also for achieving meaningful science learning. Their development and application are therefore considered fundamental objectives in science education.

3.2.4 Scientific Inquiry

Two articles focus more specifically on scientific inquiry (Rönnebeck et al., 2016; Vo & Simmie, 2025). Rönnebeck et al. (2016) describe it as engaging students in the thinking and activities of scientists. Typically, conceptualizations of scientific inquiry highlight two main dimensions: the type and scope of student activities and the degree of teacher guidance provided (e.g., Furtak et al., 2012). Furtak et al. (2012) specifically proposed a framework that distinguishes between the cognitive features of the activity—such as procedural, epistemic, and social aspects—and the level of teacher guidance.

Numerous models describe inquiry as involving student activities and competencies (e.g., Bell et al., 2010; Pedaste et al., 2015). However, they differ on which activities are central and on terminology. A particularly influential contribution is the NRC (2013) framework for K–12 education, which delineates eight interconnected practices: asking questions, developing models, planning investigations, analyzing data, using math and computation, constructing explanations, arguing from evidence, and communicating.

Other models adopt different structures. For instance, White and Frederiksen (1998) describe a progression of phases – question, predict, experiment, model, and apply; while the 5E instructional model (Bybee et al., 2000) proposes a cyclical sequence of engagement, exploration, explanation, elaboration, and evaluation. Wenning (2007) developed a detailed rubric encompassing steps such as problem identification, background research, induction and deduction, prediction, experimental design, data collection and analysis, use of statistics, and communication of results. Bell et al. (2010), in a comparative analysis of prominent models, identified nine core categories of inquiry-related activities: orienting and asking questions, generating hypotheses, planning, investigation, analysis and interpretation, modeling, drawing conclusions and evaluating, communicating, and making predictions. Pedaste et al. (2015), through a systematic review, synthesized diverse frameworks into a unified inquiry cycle consisting of five general phases—orientation, conceptualization, investigation, conclusion, and discussion.

In general, the inquiry process does not follow a strict chronological sequence; rather, it allows for multiple pathways, iterative cycles, and overlapping phases. Communication is considered an overarching skill that plays a vital role across all stages of the process (Bell et al., 2010; Pedaste et al., 2015). Building on these models, Rönnebeck et al. (2016) proposed a synthesis of activity-based conceptualizations of inquiry. Their framework identifies nine key activities, without implying a fixed sequence: identifying research questions, searching for information, formulating hypotheses and predictions, planning and conducting investigations, analyzing and evaluating data, developing explanations, constructing models, engaging in argumentation and reasoning, and communicating results. Each activity is defined with attention to both cognitive demands and epistemic function; for instance, explanations are conceived as structured arguments that link claims to evidence, and modeling includes both physical and mental representations used for prediction and reasoning.

To further organize these different models, Turner et al. (2018) grouped sixteen inquiry-related actions into three broad components: working with hypotheses (including hypothesis generation and procedure design), communication in inquiry (e.g., interpreting results and formulating questions), and hands-on inquiry (e.g., recording and visualizing data). These contributions reinforce the view of scientific inquiry as a dynamic, multi-phase process that integrates cognitive, procedural, and communicative dimensions.

Beyond these general models of scientific inquiry, specific activities like modeling and argumentation have become research areas themselves, with significant definitional variation thereby increasing the complexity of the field. (Rönnebeck et al., 2016).

3.3 Measurement Instruments and Psychometric Evaluation

3.3.1 Scientific Reasoning

The review by Opitz et al. (2017) analyzes a range of instruments developed to assess scientific reasoning, classifying them according to their theoretical foundations and shared features. Three main types of tools emerge from this analysis.

The first type is performance-based instruments, which assess scientific reasoning by involving students in concrete scientific tasks, often conducted in laboratory settings. These activities are sometimes inspired by Piagetian experiments, such as using a balance beam, pendulum tasks, or simple chemical procedures. The goal is to capture reasoning processes in action, providing richer insights than traditional written tests.

The second type involves paper-and-pencil tests, which are the most commonly used. These instruments often rely on multiple-choice items to evaluate specific components of scientific reasoning, such as the control of variables. 'Lawson's Classroom Test of Scientific Reasoning' (1978) is one of the most cited tools in the field. However, open-ended and constructed-response items are also used to assess more complex reasoning processes like explanation and argumentation.

The third type relates to the context of assessment. Opitz et al. (2017) distinguish between domain-general and domain-specific approaches. Domain-general assessments treat scientific reasoning as a general, context-free cognitive skill, measured using abstract problems. In contrast, domain-specific assessments recognize the role of prior knowledge and present problems situated in specific disciplines (e.g., biology, chemistry, physics) or real-world contexts.

In terms of psychometric properties, Opitz et al. (2017) note that early instruments (1973–1989) more frequently reported reliability analyses (10 out of 11 tests), compared to more recent ones (2002–2013), where only 17 out of 27 did so, suggesting a decline in the systematic reporting of psychometric quality over time.

Edelsbrunner and Dablander (2019) provide a more focused review of psychometric modeling in scientific reasoning research. They examine 11 empirical studies and distinguish two major approaches: the Rasch model and the broader Item Response Theory. Rasch modeling emerged as the most widely used, particularly due to its emphasis on item fit statistics, similar to practices in large-scale assessments. However, the authors highlight critical issues, especially the overextension of Rasch-based interpretations. For example, some studies misused item infit statistics to assess dimensionality, which these metrics do not validly support.

More broadly, both reviews identify a lack of detailed reporting on psychometric properties and inconsistencies between the methods used and the conclusions drawn. They call for more rigorous psychometric practices to ensure instrument reliability and validity.

3.3.2 Scientific Literacy

The review by Istyadji & Sauqina (2023) examines tools developed to assess scientific literacy. Most instruments (32 out of 46 studies focused on tool development) use multiple-choice formats. Other formats include open-ended tests and rubrics, with some tools combining various types to assess different dimensions of scientific literacy. Many instruments are grounded in the PISA framework, incorporating its competencies and knowledge domains.

The instruments typically assess cognitive knowledge (content, procedural, and epistemic), the three PISA competencies (explaining phenomena scientifically, evaluating and designing scientific inquiry, and

interpreting data and evidence), and student attitudes toward science. Some tools, particularly those for large-scale assessments aligned with PISA, are computer-based.

The review by Roy et al. (2025) adds further examples. For instance, Benjamin et al. (2017) developed the Scientific Literacy Survey for College Preparedness in STEM, which covers attitudes and behaviors, content knowledge, and reasoning skills. Most instruments are designed for secondary education, while assessing scientific literacy in preschoolers is more challenging. Among the 79 articles reviewed, 59 adopted quantitative methods, 11 used mixed methods, and 9 used qualitative designs. Data collection techniques included standardized tests (48 studies), questionnaires (20), interviews (16), and observations (11).

Istyadji & Sauqina (2023) report that Rasch models and Classical Test Theory are the most widely used psychometric approaches. Rasch models are often applied for item calibration and improving measurement precision. Reliability is typically assessed using Cronbach's Alpha, with some studies also reporting test-retest and inter-rater reliability. In Rasch-based tools, metrics like the Person Separation Index are reported. By contrast, Roy et al. (2025) do not specify the psychometric models used in the assessment of the tools they describe.

3.3.3 Science Process Skills (SPS)

According to Gizaw and Sota (2023), the tools used to measure SPS are diverse, reflecting the complexity of both cognitive and practical elements in science learning. Besides specialized tests focused on logical reasoning and experimental skills, researchers use complementary instruments such as questionnaires, interviews, reflective journals, and attitude scales. In classroom settings, observation sheets and checklists help monitor student behavior during experiments. Creativity tests and general science achievement tests are also occasionally used.

These tools aim to assess not just content knowledge but how students apply it through mental and practical activities. Practical experiments are especially valuable for observing students formulating hypotheses, designing investigations, and analyzing data. Written and oral outputs, such as reports and presentations, are important indicators of SPS development. Rubrics are often used to structure and score student performance systematically.

The meta-synthesis by Yıldırım et al. (2016) shows that questionnaires were the most common instruments in the 200 Turkish studies reviewed. Multiple-choice questions appeared in 156 studies, favored for their low cost and scalability. Open-ended questions (38 studies) and Likert scales (6 studies) were also used. Interviews (18 studies), document analysis (15), classroom observations (14), rubrics (12), and worksheets (7) appeared in smaller numbers.

Similarly, Fugarasti et al. (2019) found that paper-and-pencil tests were predominant (39% of studies), often using multiple-choice or essay formats. Observation sheets (18%), rubrics (13%), questionnaires (9%), worksheets (8%), and interview protocols (7%) were also employed, sometimes in combination.

While emphasizing the importance of reliability and validity, the studies included in the reviews provide limited detail on the psychometric properties of the instruments. However, some studies employed data triangulation – such as combining multiple-choice and open-ended items with interviews or observations – to enhance reliability.

3.3.4 Scientific Inquiry

The review by Rönnebeck et al. (2016) focuses on how scientific inquiry activities are implemented and assessed. To evaluate students' modeling skills, paper-and-pencil tests with multiple-choice, constructed-response, or open-ended items are frequently used. Explanatory skills are typically assessed through discourse analysis and examination of written responses. For argumentation and reasoning, the Toulmin (1958) model and its adaptations are widely applied. Kaberman and Dori (2009) offer a detailed taxonomy to analyze hypotheses based on content, cognitive level, and domain-specific understanding.

Vo & Simmie (2025) observe that the impact of assessment format on evaluating student inquiry skills is rarely addressed. They identify three main approaches, noting a growing trend toward digital environments. Hands-on performance assessments, used in 17.6% of studies, provide authentic context evaluations. However, the most common format is traditional test batteries (41.1%), including multiple-choice, short-answer, and essay questions. Digital test batteries (39.7%) and simulation-based assessments (37.0%) are increasingly common. Multiple-choice items appear in 71.4% of studies, and open-ended formats in 69.8%. Nearly half (44.5%) of studies combine both formats.

Neither review offers detailed information about the psychometric validation of these instruments.

3.4 Evidence of Validity

3.4.1 Scientific Reasoning

Opitz et al. (2017) provide an overview of various forms of validity considered in the instruments they reviewed. These include: content validity, referring to the extent to which test items adequately represent the intended construct; construct validity, which assesses whether the instrument accurately measures scientific reasoning and correlates appropriately with other instruments targeting similar constructs; concurrent and divergent validity, established through comparisons with other scientific reasoning tests or with general cognitive assessments such as IQ tests; criterion validity, which examines whether test scores predict relevant external outcomes or variables.

3.4.2 Scientific Literacy

Istyadji & Sauqina (2023) stress the importance of thoroughly testing the validity and reliability of scientific literacy instruments to ensure their appropriateness for educational or research settings. Content validity is often assessed through expert judgment by specialists in science education, chemistry, or linguistics, ensuring that the items reflect the intended construct; construct validity is commonly established using statistical techniques, particularly Confirmatory Factor Analysis and Exploratory Factor Analysis; face validity, the

clarity, structure, and appearance of the instrument, is also considered to ensure that the test is understandable and usable.

3.4.3 Science Process Skills (SPS)

The meta-synthesis by Yıldırım et al. (2016) does not provide detailed information on the specific types of validity reported in the studies reviewed. However, the authors recommend methodological strategies such as employing multiple data collection tools and validating instruments with large or diverse samples. These recommendations implicitly aim to strengthen construct and content validity. Similarly, the reviews by Fugarasti et al. (2019) and Gizaw and Sota (2023) highlight the importance of using valid and reliable tools for assessing SPS, though they do not discuss specific validation procedures. Overall, while the studies emphasize the need for validity, they provide minimal information on how it was formally tested.

3.4.4 Scientific Inquiry

Rönnebeck et al. (2016) raise significant concerns about the construct validity of scientific inquiry assessments. They point out that many studies lack clear definitions of the inquiry construct, which makes it difficult to compare results and may obscure the effects of different instructional approaches. Additionally, inconsistent application of theoretical concepts across studies undermines the coherence and validity of findings. Although the review does not explicitly describe formal validation procedures, it notes that many studies provide limited or unclear information about the theoretical foundations of their instruments. Vo & Simmie (2025) similarly do not report the use of specific statistical validation procedures, further highlighting the lack of robust validity evidence in this area.

Table 2 provides a summary of the results in relation to the research questions, organized according to the three phases of construct validation.

Table 2 – *Summary of Findings*[b]

Construct	Reviews	Substantive phase	Structural phase		External phase
		Conceptualization	Measurement Instruments	Psychometric Evaluation	Validity Evidence
Scientific Reasoning	Edelsbrunner & Dablander, 2019; Opitz et al., 2017	Cognitive skills related to hypothesis generation, testing, and knowledge reflection; foundational for scientific literacy.	Performance-based tasks, paper-and-pencil tests; both domain-general and domain-specific contexts.	Reliability, Rasch model, Item Response Theory.	Content, construct, concurrent, divergent and criterion validity.
Scientific Literacy	Istyadji & Sauqina, 2023; Roy et al., 2025	Multidimensional: includes conceptual understanding, inquiry skills, epistemological awareness, and attitudes.	Multiple-choice tests; PISA-based frameworks; standardized assessments.	Rasch model, Classical Test Theory; Cronbach's alpha, test-retest, inter-rater reliability, separation Index.	Content, construct and face validity.
Science Process Skills (SPS)	Fugarasti et al., 2019; Gizaw & Sota, 2023; Yıldırım et al., 2016	Core investigative skills, organized hierarchically as basic and integrated process skills.	Written and practical assessments; observational tools; creativity and general achievement tests.	Limited data; some studies used triangulation to support reliability.	N/A
Scientific Inquiry	Rönnebeck et al., 2016; Vo & Simmie, 2025	A dynamic, non-linear process integrating cognitive, procedural, and communicative dimensions; varies by activity type and teacher guidance.	Written tasks, performance-based assessments, digital tools.	N/A	N/A

[b]*Note*: N/A = Not Applicable.

4. Discussion

This umbrella review aimed to systematically map how scientific reasoning is defined, measured and validated in existing systematic reviews and meta-syntheses. Although the focus was primarily on scientific reasoning as the core construct, reviews addressing closely related concepts such as scientific enquiry, science process skills (SPS) and scientific literacy were also included. This methodological approach was based on previous landmark reviews of scientific reasoning (e.g. Edelsbrunner & Dablander, 2019; Opitz et al., 2017), which also used broad search strategies. These studies acknowledged the conceptual overlap of the constructs. Examining these constructs together enables this review to offer a more comprehensive understanding of how scientific reasoning is situated within broader scientific competencies. Ultimately, nine systematic reviews and meta-syntheses published between 2016 and 2025 met the inclusion criteria and formed the basis of this synthesis.

Our findings reveal an absence of a shared conceptual framework that connects the four constructs (scientific reasoning, scientific inquiry, scientific literacy, and science process skills) in a coherent and theoretically grounded way. Although they are all widely used in educational research and policy, they are rarely discussed within a unified theoretical model. Based on our synthesis, we propose the hypothesis that these constructs are not equivalent. Specifically, scientific inquiry appears to operate primarily as a set of pedagogical activities or instructional approaches, whereas scientific reasoning, scientific literacy and SPS are more clearly defined as skills or cognitive abilities. This implies that the constructs do not operate on the same logical level. This conceptual ambiguity, coupled with the lack of an overarching framework that clearly links these four constructs, results in a fragmented and partially redundant field. This hinders the accumulation of knowledge, as researchers may investigate the same phenomena under different labels unknowingly, or conflate distinct constructs. This leads to both conceptual overlap and theoretical disintegration (Anvari et al., 2025).

Based on the literature reviewed, we tentatively outline a possible framework that may help to conceptualize the interplay between the four constructs. Scientific literacy could be viewed as the overarching educational objective, widely recognized in international discourse as essential for equipping citizens to engage critically and constructively with science in society (Roy et al., 2025). Whether framed in terms of content mastery (Vision I) or real-world application (Vision II), scientific literacy encompasses an understanding of, ability to use and appreciation for scientific concepts. Scientific enquiry may be the primary process through which scientific literacy develops. This involves engaging learners in authentic scientific practices and ways of thinking, such as posing questions, designing investigations, analyzing data and reasoning from evidence (Rönnebeck et al., 2016). Within this process, science process skills (SPS) could be considered the foundational abilities — ranging from basic skills such as observing and predicting, to more complex skills such as experimenting and modelling — that enable students to carry out inquiry effectively (Yıldırım et al., 2016; Gizaw & Sota, 2023). In turn, scientific reasoning could be seen as the core cognitive capacity activated and refined through inquiry and the application of SPS. It encompasses the mental processes, such as generating hypotheses, controlling variables and evaluating evidence, that give scientific inquiry meaning. In this view, scientific reasoning may function as the 'mental engine' that drives the practical use of SPS within inquiry-

based learning, thereby supporting the development of scientific literacy. In summary, this hypothetical framework suggests that scientific literacy is the ultimate goal, scientific enquiry is the developmental pathway towards it, SPS are the operational tools, and scientific reasoning is the cognitive foundation that enables their effective use.

We structured our analysis around the three validation phases proposed in the literature on construct validation: (1) the substantive phase focused on how the constructs were defined and conceptualized; (2) the structural phase investigated the measurement instruments used, their characteristics and their reported psychometric properties; (3) the external phase assessed the presence of validity evidence relating test scores to external criteria or constructs. While the substantive phase provided a relatively clear picture of theoretical positions and models, the findings from the structural phase, and more specifically from the external phase, were lacking. Most reviews provided little to no detail about the measurement instruments, such as the number of items, response format, scoring models, characteristics of the sample, intended performance level (typical versus maximal) and the psychometric models used to assess structure and reliability. Moreover, evidence of external validity was almost entirely absent. This lack of information can be attributed to two main factors. Firstly, none of the included reviews explicitly aimed to systematically catalogue measurement instruments in terms of psychometric properties. Secondly, as noted in the discussions of several of the reviews, many primary studies failed to report psychometric details. Consequently, the lack of sufficient reporting prevented a clear assessment of their empirical implementation.

All reviews expressed concern about the limited attention given to psychometric rigor in the original studies analyzed. Specifically, they highlighted the scarcity of information relating to instrument structure, reliability, and external validation. These issues are consistent with broader problems in psychological research, where a lack of robust validation practices can undermine construct clarity and the quality of measurements (Borsboom, 2006; Flake et al., 2017). Robust and validated measurement instruments are essential for both scientific progress and evidence-based practice in education. Without such tools, researchers cannot compare findings across studies, educators cannot reliably assess student learning and policymakers lack a solid foundation on which to develop curricula. Recent literature has emphasized that measurement instruments must be designed, tested and refined through ongoing validation processes rather than being treated as static, one-time products (Flake et al., 2017).

An additional methodological consideration concerns the degree of overlap among the primary studies across reviews. The CCA value of 1.76% falls within the “slight overlap” range, indicating that the reviews synthesized largely distinct sets of evidence with minimal redundancy. This low overlap enhances the breadth of the umbrella review, but it also reflects the fragmentation of the field, where related constructs are often examined in isolation. At the same time, the fact that not all reviews provided complete lists of primary studies limits the comprehensiveness of the overlap analysis, an issue that is addressed further in the limitations section.

4.1 Limitations and directions for future research

This umbrella review has several limitations. Firstly, we did not focus on the specific populations or educational levels to which these constructs and instruments were applied. This was partly due to the limited detail in the included reviews and the conceptual heterogeneity observed, which made focusing on definitions and measurement approaches more meaningful. Secondly, a limitation concerns the calculation of overlap among reviews. The CCA could only be computed for the seven reviews that provided a list of primary studies, whereas two (Gizaw & Sota, 2023; Roy et al., 2025) did not make this information available. This absence of complete data prevents a fully comprehensive estimate and may conceal additional redundancies. Thirdly, another limitation relates to the grading of the overall strength of the evidence, a step often expected in umbrella reviews. In this case, the synthesis included only nine reviews, each addressing partly different constructs (scientific reasoning, scientific inquiry, science process skills, and scientific literacy). This heterogeneity makes it difficult, and perhaps misleading, to attempt a quantitative grading across studies. For this reason, we relied on a narrative appraisal that considered three dimensions: (a) the consistency of findings across reviews, (b) the methodological quality of the syntheses themselves, and (c) the breadth and robustness of the underlying evidence. From this perspective, evidence relating to the substantive phase of validation appears to be emerging but conceptually heterogeneous; findings on structural validation suggest a moderate and relatively consistent level of support; whereas external validation seems weaker, reflecting the limited number of systematic investigations available. These characterizations are qualitative and intended to offer orientation rather than definitive grading. Our narrative weighting of review quality and appraisal of evidence strength should therefore be understood as a pragmatic strategy to enhance transparency, while recognizing both the heterogeneity of the evidence base and the exploratory nature of this synthesis.

4.2 Directions for future research

Several key areas for future research emerge from this synthesis. (1) Clarifying construct interrelationships and developing unified models: a critical priority is addressing the conceptual ambiguity that pervades this field. Researchers must develop and test a unified model that clarifies the relationships between the four constructs (scientific reasoning, scientific literacy, SPS, and scientific inquiry) and translates them into operational measurement criteria. Such a model should integrate the different aspects already present in international frameworks (e.g., scientific literacy, inquiry, and SPS) and be empirically tested, for example in the context of PISA, which offers both wide coverage and increasing attention to transversal competencies. (2) Improving methodological rigor: the field requires greater adherence to contemporary views of validity, emphasizing a comprehensive and systematic approach to gathering evidence (Flake et al., 2017). Failure to do so risks undermining cumulative science (Anvari et al., 2025). (3) Systematic mapping of instruments: none of the reviews specifically catalogued and evaluated existing instruments. A dedicated systematic review is needed to provide a detailed inventory, including design, intended population, development process, and psychometric validation. (4) Expanding populations and contexts: most studies to date have been conducted with secondary school students in formal educational settings within technologically advanced countries.

Future research should extend to non-Western populations and adult learners, in higher education or professional training, to capture developmental trajectories of scientific reasoning across the lifespan.

In light of these findings, it is important to recognize that these constructs — scientific reasoning, scientific enquiry, scientific literacy and science process skills — are theoretically relevant and also core educational objectives worldwide. International frameworks and assessments, such as the OECD's Programme for International Student Assessment (PISA), the National Research Council in the US, and multiple national curricula, emphasize the development of these skills as being essential for informed citizenship and lifelong learning. Therefore, clear construct definitions and rigorous measurement instrument validation are prerequisites for informed educational decision-making, not academic luxuries. For these constructs to guide teaching, assessment and policy, they must be measurable in a conceptually coherent and psychometrically sound way. Only then can they fulfil their potential to shape meaningful learning outcomes and support evidence-based educational reform.

5. Conclusion

This umbrella review consolidates and critically evaluates how scientific reasoning and related constructs have been conceptualized, measured, and validated in the existing review literature. By integrating nine systematic reviews and meta-syntheses, it highlights the absence of a unified conceptual framework linking scientific reasoning, scientific inquiry, science process skills, and scientific literacy, and reveals major gaps in the reporting and validation of measurement instruments.

From a methodological standpoint, we addressed key features expected in umbrella reviews by quantifying overlap among primary studies, narratively weighting conclusions by review quality, and providing a qualitative grading of the strength of evidence. Although these procedures cannot substitute for more standardized approaches—which are still lacking in the educational and psychological sciences—they enhance the transparency and interpretability of the synthesis.

Taken together, the findings underscore the urgent need for clearer construct definitions, unified theoretical models, and greater psychometric rigor in instrument development. Such advances are not only of academic interest but are also essential to support evidence-based educational policies and practices. By clarifying what is meant by scientific reasoning and how it can be reliably assessed, future research can better inform teaching, learning, and large-scale assessments, ultimately strengthening the preparation of scientifically literate citizens.

6. Declarations

Funding. No funding was received for conducting this study.

Competing interests. The authors have no relevant financial or non-financial interests to disclose.

Ethical approval. This article is based on a review of published literature. No studies with human participants or animals were conducted by the authors. Therefore, ethical approval and informed consent were not required.

Preregistration and Data Availability. This study was preregistered on the Open Science Framework (OSF) [https://osf.io/jvye7]. The preregistration file includes the full protocol detailing the objectives, inclusion and exclusion criteria, search strategy, study selection process, and data extraction procedures. In addition, two supplementary Excel file is openly available on OSF, including: (a) the list of records retrieved and screened; (b) documentation of inclusion/exclusion decisions; (c) the coding and assessment sheet used for evaluating methodological quality; (d) the extraction template with the information collected from each review; and (e) the Corrected Covered Area (CCA) matrix with the calculation of overlap among reviews.

CRediT author statement

Rossella Caliciuri: Conceptualization, Methodology, Formal analysis, Investigation, Data curation, Writing – original draft, Writing - Review & Editing, Project administration.

Giulia Moncalieri: Formal analysis, Writing – review & editing.

Margherita Lanz: Conceptualization, Supervision, Writing – review & editing.

All authors approved the final version of the manuscript.

Bibliography

Abd Rauf, R. A., Rasul, M. S., Mans, A. N., Othman, Z., & Lynd, N. (2013). Inculcation of science process skills in a science classroom. *Asian Social Science*, *9*(8), 1911-2017. https://doi.org/10.5539/ass.v9n8p47

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing*. Washington, DC: Joint Committee on Standards for Educational and Psychological Testing.

Anvari, F., Alsalti, T., Oehler, L., Hussey, I., Elson, M., & Arslan, R. C. (2024). A fragmented field: Construct and measure proliferation in psychology. *Preprint]. https://osf. io/preprints/psyarxiv/b4muj*.

Aromataris E, Fernandez RS, Godfrey C, Holly C, Khalil H, Tungpunkom P. Umbrella reviews. In: Aromataris E, Lockwood C, Porritt K, Pilla B, Jordan Z, eds. *JBI Manual for Evidence Synthesis*. Adelaide: JBI; 2024. https://synthesismanual.jbi.global

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), 010115. https://doi.org/10.1103/PhysRevPhysEducRes.18.010115

Bell, T., Urhahne, D., Schanze, S., & Ploetzner, R. (2010). Collaborative inquiry learning: Models, tools, and challenges. *International journal of science education*, *32*(3), 349-377. https://doi.org/10.1080/09500690802582241

Benjamin, T. E., Marks, B., Demetrikopoulos, M. K., Rose, J., Pollard, E., Thomas, A., & Muldrow, L. L. (2017). Development and validation of scientific literacy scale for college preparedness in stem with freshmen from diverse institutions. *International Journal of Science and Mathematics Education, 15*(4), 607–623. https://doi.org/10.1007/s10763-015-9710-x

Borsboom, D. (2006). The attack of the psychometricians. *Psychometrika*, *71*(3), 425-440. https://doi.org/10.1007/s11336-006-1447-6

Bybee, R. W. (2000). Teaching science as inquiry. In J. Minstrell & E. H. van Zee (Eds.), *Inquiring into inquiry learning and teaching in science* (pp. 20–46). American Association for the Advancement of Science.

Chakraborty, D., & Kidman, G. (2022). Inquiry process skills in primary science textbooks: Authors and publishers' intentions. *Research in Science Education*, *52*(5), 1419-1433. https://doi.org/10.1007/s11165-021-09996-4

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and psychological measurement*, *20*(1), 37-46. https://doi.org/10.1177/001316446002000104

Coil, D., Wenderoth, M. P., Cunningham, M., & Dirks, C. (2010). Teaching the process of science: faculty perceptions and an effective methodology. *CBE—Life Sciences Education*, *9*(4), 524-535. https://doi.org/10.1187/cbe.10-01-0005

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin, 52*(4), 281–302. https://doi.org/10.1037/h0040957

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

Edelsbrunner, P. A., & Dablander, F. (2019). The psychometric modeling of scientific reasoning: A review and recommendations for future avenues. *Educational Psychology Review*, *31*, 1-34. https://doi.org/10.1007/s10648-018-9455-5

Engelmann, K., Neuhaus, B. J., & Fischer, F. (2016). Fostering scientific reasoning in education–meta-analytic evidence from intervention studies. *Educational research and evaluation*, *22*(5-6), 333-349. https://doi.org/10.1080/13803611.2016.1240089

European Commission and Directorate-General for Research and Innovation. (2015). *Science education for responsible citizenship: Report to the European Commission of the expert group on science education*. Publications Office. https://doi.org/10.2777/12626

Fatimah, S., Sarwi, S., Linuwih, S., & Dewi, N. R. (2025). Facilitating Students' Scientific Reasoning Skills and Sustainability Literacy in Higher Education-Needs Analysis and Suggested Model. *International Journal of Scientific Multidisciplinary Research*, *3*(5), 691-710. https://doi.org/10.55927/ijsmr.v3i5.261

Finley, F. N. (1983). Science processes. *Journal of research in science teaching*, *20*(1), 47-54. https://doi.org/10.1002/tea.3660200105

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., Pek, J., & Hehman, E. (2017). Construct validation in social and personality research: Current practice and recommendations. *Social Psychological and Personality Science, 8*(4), 370–378. https://doi.org/10.1177/1948550617693063

Fugarasti, H., Ramli, M., & Muzzazinah. (2019, December). Undergraduate students' science process skills: A systematic review. In *AIP Conference Proceedings* (Vol. 2194, No. 1, p. 020030). AIP Publishing LLC. https://doi.org/10.1063/1.5139762

Furtak, E. M., Seidel, T., Iverson, H., & Briggs, D. C. (2012). Experimental and quasi-experimental studies of inquiry-based science teaching: A meta-analysis. *Review of educational research*, *82*(3), 300-329. https://doi.org/10.3102/0034654312457206

Gizaw, G., & Sota, S. (2023). Improving science process skills of students: A review of literature. *Science Education International*, *34*(3), 216-224. https://doi.org/10.33828/sei.v34.i3.5

Gough, D., & Richardson, M. (2018). Systematic reviews. In *Advanced research methods for applied psychology* (pp. 63-75). Routledge.

Holbrook, J., & Rannikmae, M. (2009). The meaning of scientific literacy. *International journal of environmental and science education*, *4*(3), 275-288.

Hurd, P. D. (1958). Science literacy: Its meaning for American schools. *Educational leadership*, *16*(1), 13-16.

Istyadji, M. & Sauqina (2023). Conception of scientific literacy in the development of scientific literacy assessment tools: A systematic theoretical review. *Journal of Turkish Science Education*, *20*(2), 281-308. https://doi.org/10.36681/tused.2023.016

Kaberman, Z., & Dori, Y. J. (2009). Question posing, inquiry, and modeling skills of chemistry students in the case-based computerized laboratory environment. *International Journal of Science and Mathematics Education, 7*, 597–625. https://doi.org/10.1007/s10763-007-9118-3

Klahr, D., & Dunbar, K. (1988). Dual space search during scientific reasoning. *Cognitive science*, *12*(1), 1-48. https://doi.org/10.1016/0364-0213(88)90007-9

Kuhn, D. (1989). Children and adults as intuitive scientists. *Psychological review*, *96*(4), 674. https://doi.org/10.1037/0033-295X.96.4.674

Kuhn, D. (1991). *The skills of argument*. Cambridge University Press. https://doi.org/10.1017/CBO9780511571350

Laugksch, R. C. (2000). Scientific literacy: A conceptual overview. *Science education*, *84*(1), 71-94. https://doi.org/10.1002/(SICI)1098-237X(200001)84:1<71::AID-SCE6>3.0.CO;2-C

Lawson, A. E. (1978). The development and validation of a classroom test of formal reasoning. *Journal of Research in Science Teaching*, 15, 11–24. https://doi.org/10.1002/tea.3660150103

Lawson, A. E. (1995). *Science teaching and the development of thinking*. Belmont, CA: Wadsworth.

Loevinger, J. (1957). Objective tests as instruments of psychological theory. *Psychological reports*, *3*(3), 635-694.

Millar, R. (1997). Science education for democracy: What can the school curriculum achieve? In *Science today: problem or crisis?* (pp. 87–101). Routledge.

Millar, R. (2006). Twenty First Century Science: Insights from the design and implementation of a scientific literacy approach in school science. *International Journal of Science Education*, *28*(13),1499–1521. https://doi.org/10.1080/09500690600718344

Miller, J. D. (1983). Scientific literacy: A conceptual and empirical review. *Daedalus*, 29-48.

National Research Council. 2012. *A Framework for K-12 Science Education: Practices, Crosscutting Concepts, and Core Ideas.* Washington, DC: The National Academies Press. https://doi.org/10.17226/13165

Opitz, A., Heene, M., & Fischer, F. (2017). Measuring scientific reasoning–a review of test instruments. *Educational Research and Evaluation*, *23*(3-4), 78-101. http://dx.doi.org/10.1080/13803611.2017.1338586

Organisation for Economic Cooperation and Development [OECD]. (2023). *PISA 2025 Science Framework* (Second draft) [PDF]. OECD Publishing. https://pisa-framework.oecd.org/science-2025/assets/docs/PISA_2025_Science_Framework.pdf

Osborne, J. (2010). Arguing to learn in science: The role of collaborative, critical discourse. *science*, *328*(5977), 463-466. https://doi.org/10.1126/science.1183944

Özgelen, S. (2012). Students' science process skills within a cognitive domain framework. *Eurasia Journal of Mathematics, Science and Technology Education*, *8*(4), 283-292. https://doi.org/10.12973/eurasia.2012.846a

Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., Shamseer, L., Tetzlaff, J. M., Akl, E. A., Brennan, S. E., Chou, R., Glanville, J., Grimshaw, J. M., Hróbjartsson, A., Lalu, M. M., Li, T., Loder, E. W., Mayo-Wilson, E., McDonald, S., … Moher, D. (2021). The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. *BMJ (Clinical Research Ed.)*, n71. https://doi.org/10.1136/bmj.n71

Pedaste, M., Mäeots, M., Siiman, L. A., De Jong, T., Van Riesen, S. A., Kamp, E. T., ... & Tsourlidaki, E. (2015). Phases of inquiry-based learning: Definitions and the inquiry cycle. *Educational research review*, *14*, 47-61. https://doi.org/10.1016/j.edurev.2015.02.003

Piaget, J., & Inhelder, B. (1958). *The growth of logical thinking from childhood to adolescence: An essay on the construction of formal operational structures*. Abingdon, Oxon: Routledge.

Pieper, D., Antoine, S. L., Mathes, T., Neugebauer, E. A. M., & Eikermann, M. (2014). *Systematic review finds overlapping reviews were not mentioned in every other overview*. Journal of Clinical Epidemiology, 67(4), 368–375. https://doi.org/10.1016/j.jclinepi.2013.11.007

Roberts, D. A., & Bybee, R. W. (2014). Scientific literacy, science literacy, and science education. In N. G. Lederman & S. K. Abell (Eds.), *Handbook of research on science education* (Vol. 2, pp. 559–572). Routledge.

Rönnebeck, S., Bernholt, S., & Ropohl, M. (2016). Searching for a common ground – A literature review of empirical research on scientific inquiry activities. *Studies in Science Education*, *52*(2), 161–197. https://doi.org/10.1080/03057267.2016.1206351

Roy, G., Sikder, S., & Danaia, L. (2025). Adopting scientific literacy in early years from empirical studies on formal education: a systematic review of the literature. *International Journal of STEM Education*, *12*(1), 1-24. https://doi.org/10.1186/s40594-025-00547-1

Schwartz, R. S., Lederman, J. S., & Enderle, P. J. (2023). Scientific inquiry literacy: The missing link on the continuum from science literacy to scientific literacy. In N. G. Lederman, D. L. Zeidler, & J. S. Lederman (Eds.), *Handbook of research on science education* (pp. 749–782). Routledge.

Shamos, M. H. (1995). *The myth of scientific literacy*. Rutgers University Press.

Toulmin, S. E. (1958). *The uses of argument*. Cambridge, UK: Cambridge University Press.

Turner, R. C., Keiffer, E. A., & Salamo, G. J. (2018). Observing inquiry-based learning environments using the scholastic inquiry observation instrument. *International Journal of Science and Mathematics Education*, *16*(8), 1455-1478. https://doi.org/10.1007/s10763-017-9843-1

Vo, D. V., & Simmie, G. M. (2025). Assessing scientific inquiry: A systematic literature review of tasks, tools and techniques. *International Journal of Science and Mathematics Education*, *23*(4), 871-906. https://doi.org/10.1007/s10763-024-10498-8

Wenning, C. J. (2007). Assessing inquiry skills as a component of scientific literacy. *Journal of Physics Teacher Education Online*, *4*(2), 21-24.

White, B. Y., & Frederiksen, J. R. (1998). Inquiry, modeling, and metacognition: Making science accessible to all students. *Cognition and instruction*, *16*(1), 3-118. https://doi.org/10.1207/s1532690xci1601_2

Yildirim, M., Çalik, M., & Özmen, H. (2016). A Meta-Synthesis of Turkish Studies in Science Process Skills. *International Journal of Environmental and Science Education, 11*(14), 6518-6539.

Zimmerman, C., & Klahr, D. (2018). Development of scientific thinking. *Stevens' handbook of experimental psychology and cognitive neuroscience*, *4*, 1-25. https://doi.org/10.1002/9781119170174.epcn407

Chapter 4 - Scientific Reasoning Scale in Italy: Validation Studies[2]

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.

Keywords: scientific reasoning, classical test theory, confirmatory factor analysis, unified view of validity, contemporary view of validity, measurement invariance, structural equation modeling

[2] *This chapter was co-authored by Rossella Caliciuri and Margherita Lanz and corresponds to an article published in Thinking Skills and Creativity (Caliciuri & Lanz, 2026).*

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 observation and evaluation of consequences (Bao et al., 2009; Bao et al., 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, 2003), 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' (AERA, APA, & 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; Zumbo, 2005; Sorgente & Zumbo, 2025). 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 & 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. Figure 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.

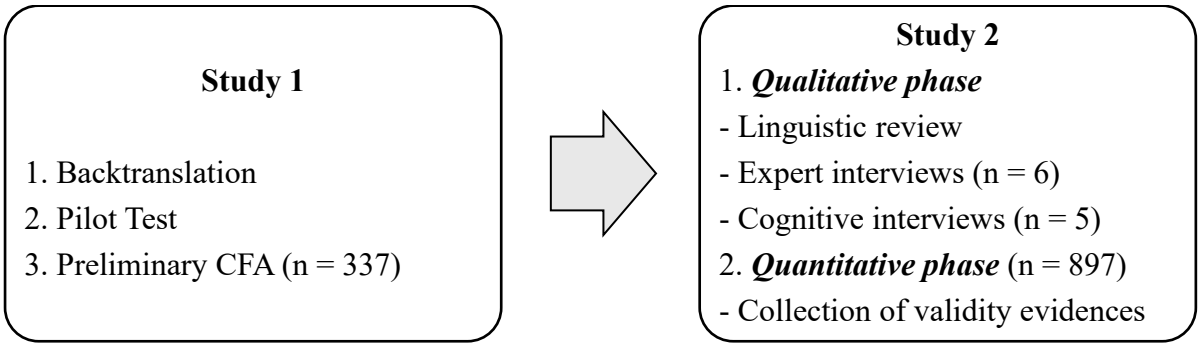

Fig. 1 *Structure of the multi-study and multi-method project conducted to provide validity evidence for the Scientific Reasoning Scale (SRS)*

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 & 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 ≥ .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 '5.3 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 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 [blinded] 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 minutes 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 version 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, Morsanyi, Chiesi, Donati, & Hamilton, 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, West, & Stanovich, 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 = .04 (.02 - .06); CFI = .996; WRMR = .80]. The factor loadings were all high (> |.71|) and significant (p < .001). Internal consistency score was good (McDonald's ω = .80).

The Probabilistic Reasoning Scale (PRS; Primi, Morsanyi, Galli, & Chiesi, 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 = .05 (.04 - .05); CFI = .954; WRMR = 1.34]. The factor loadings were all high (> |.30|) and significant (p < .001). Internal consistency score was good (McDonald's ω = .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 = .05 (.04 - .05); CFI = .925; SRMR = .05]. The factor loadings were all high (> |.77|) and significant (p < .001). Internal consistency scores ranged from acceptable to excellent, with McDonald's ω = .73 (MEDB), .81 (EEB), .85 (PSIB), .91 (SB), and .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 $\boldsymbol{\rho}$ = .476).

Climate Change Awareness: individuals' concern for climate change issues will be measured through six items used in Robba et al., (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 = .06 (.04 - .08), p = .278; CFI = .986; SRMR = .17]. The factor loadings were all high (> |.79|) and significant (p < .001). Internal consistency score was excellent (McDonald's ω = .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 and 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 .08 considered reasonable and those below .05 considered good (Lai & Green, 2016; Kline, 2023). CFI and TLI values close to 1 suggest a good model fit, whereas values below .90 indicate a poor fit (Lai & Green, 2016). SRMR values close to zero reflect a better fit, with values below .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 .010 or more (Cheung & Rensvold, 2002; Meade, Johnson, & Braddy, 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, Balla, & McDonald, 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 .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).

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%

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 < .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 = .029 (.017 .041), p = .999; CFI = .966; WRMR = .953]. As shown in Figure 2, all factor loadings were high (< .3) and significant (p < .001), all ten items load significantly on a single latent factor representing scientific reasoning, with standardized loadings ranging from .336 to .640. These results confirm the unidimensionality of the scale. Composite reliability was estimated for SRS: ω = .612.

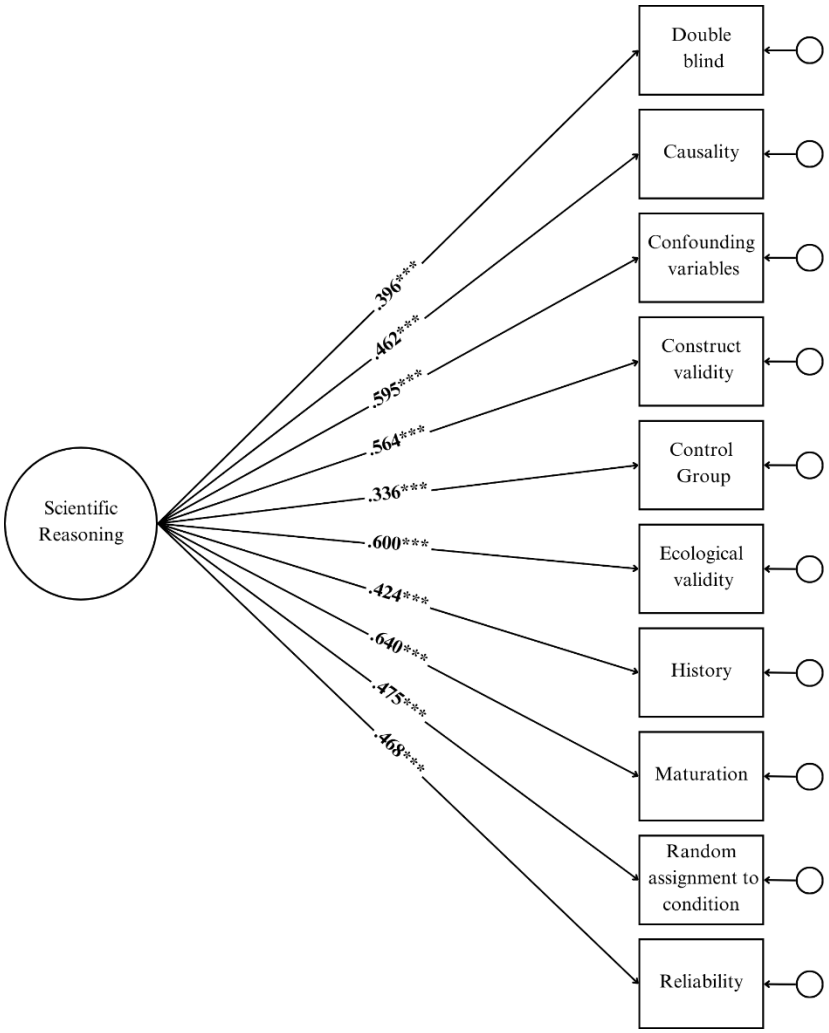

Fig. 2 *Confirmed factorial structure of Scientific Reasoning Scale (SRS)*[a]

[a]*Note. Numbers on the arrows indicate standardized factor loadings (range: .336 to .640). The items significantly load onto a single latent factor representing scientific reasoning (*** indicates p < .001).*

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, Hofer & Lawrence, 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 achieved, as indicated by a significant difference between the strict and factor invariance models (Difftest $\chi^2 < .001$; $\Delta CFI = –.033$). In this case, participants with a university degree showed significantly higher factor means than those with only a high school diploma.

Table 3 *Measurement Invariance of the SRS across different demographic and ideological conditions*

Invariance	χ^2	Df	p	RMSEA (CI)	p RMSEA <= .05	CFI	WRMR	DIFF TESTχ^2	Df	p	ΔCFI
Gender (438 male vs 453 female)											
Configural	85.40	70	.102	.02 (.00 .04)	1.00	.980	1.12				
Weak	91.03	79	.167	.02 (.00 .03)	1.00	.984	1.22	8.53	9	.482	.004
Strong	114.55	88	.030	.03 (.01 .04)	1.00	.965	1.37	25.83	9	.002	-.019
freeing item 9	103.44	87	.110	.02 (.00 .04)	1.00	.978	1.30	13.44	8	.098	-.006
Strict								10.15	10	.428	
Factor Variance	102.11	88	.144	.02 (.00 .03)	1.00	.981	1.31	0.46	1	.496	.003
Factor Mean	104.50	89	.125	.02 (.00 .03)	1.00	.979	1.34	2.81	1	.094	-.002
Age (180 GenZ vs 315 GenY vs 401 GenX)											
Configural	130.90	105	.044	.03 (.01 .04)	.99	.968	1.39				
Weak	141.48	123	.122	.02 (.00 .04)	1.00	.977	1.54	16.26	18	.574	.009
Strong	163.21	141	.097	.02 (.00 .04)	1.00	.972	1.65	22.62	18	.206	-.005
Strict								25.27	20	.191	
Factor Variance	187.91	143	.007	.03 (.02 .04)	.99	.944	1.83	11.39	2	.003	-.028
Factor Mean	189.32	145	.008	.03 (.02 .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 (.00 .04)	1.00	.984	1.10				
Weak	90.48	79	.178	.02 (.00 .03)	1.00	.986	1.21	9.61	9	.383	.002
Strong	107.18	88	.081	.02 (.00 .04)	1.00	.976	1.32	18.14	9	.034	-.010
Strict								19.52	10	.034	
Factor Variance	134.88	89	<.001	.03 (.02 .05)	.99	.943	1.53	17.03	2	<.001	-.033
Employment status (134 students vs 588 workers vs 173 'other')											
Configural	126.79	105	.073	.03 (.00 .04)	1.00	.971					
Weak	136.39	123	.193	.02 (.00 .04)	1.00	.982	1.52	14.87	18	.671	.011
Strong	158.95	141	.143	.02 (.00 .04)	1.00	.976	1.64	24.10	18	.152	-.006
Strict								25.09	20	.198	
Factor Variance	171.18	143	.054	.03 (.00 .04)	1.00	.963	1.74	7.52	2	.023	-.013
Factor Mean	172.49	145	.059	.03 (.00 .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 (.01 .05)	.988	.960	1.42				
Weak	144.65	123	.089	.02 (.00 .04)	.999	.971	1.57	15.99	18	.594	.011
Strong	163.06	141	.099	.02 (.00 .04)	1.00	.971	1.66	18.55	18	.420	.000
Strict								23.27	20	.276	
Factor Variance	162.40	143	.128	.02 (.00 .04)	1.00	.974	1.68	2.00	2	.368	.003

Factor Mean	181.44	145	.022	.03 (.01 .04)	.998	.952	1.81	11.17	2	.004	-.022
Religious affiliation (474 'yes' vs 423 'no')											
Configural	97.13	70	.018	.03 (.01 .04)	.996	.966	1.20				
Weak	111.28	79	.010	.03 (.02 .04)	.997	.959	1.36	14.80	9	.097	-.007
Strong	117.89	88	.018	.03 (.01 .04)	.999	.962	1.40	6.65	9	.674	-.003
Strict								12.29	10	.266	
Factor Variance	118.52	89	.020	.03 (.01 .04)	.999	.962	1.42	1.33	1	0,2483	.000
Factor Mean	132.00	90	.003	.03 (.02 .04)	.996	.947	1.52	7.00	1	0,0082	-.015

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 .523 ($p < .001$) with the CRT-Long and .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 = .027 (.017 .037), p = 1; CFI = .973; WRMR = .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 = .017 (.004 .026), p = 1; CFI = .989; WRMR = .750]. Negative correlations were found between the SRS and all five paranormal health 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 ($r = .118$, $p < .01$) and CCB ($r = .131$, $p < .01$).

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 (>.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 and 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 (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.

6. Declarations

Ethics approval. The study protocol was approved by the Ethical Committee of the Department of Psychology of [blinded] 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.

Funding. No funding was received for conducting this 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 plan, codebook, and 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.

CRediT author statement

Rossella Caliciuri: Conceptualization, Formal analysis, Investigation, Methodology, Project administration, Resources, Visualization, Writing – original draft.

Margherita Lanz: Conceptualization, Methodology, Project administration, Resources, Supervision, Validation, Writing – review & editing.

All authors approved the final version of the manuscript.

Appendix A - Scientific Reasoning Scale – Italian adaptation

Scientific Reasoning Scale – Italian adaptation

Le domande che seguono valutano le abilità di ragionamento scientifico. Ogni scenario presenta un esperimento o una situazione di ricerca, e per ciascuno bisognerà scegliere l'opzione che riflette la corretta applicazione dei principi scientifici. Per rispondere non è necessaria alcuna conoscenza specifica pregressa. La risposta corretta dipende esclusivamente dalle informazioni fornite nel testo.

Concept	Item
Cieco/doppio cieco	In un esperimento per valutare le preferenze di gusto del caffè, un ricercatore mette il caffè del marchio A e il caffè del marchio B in tazze identiche, per distinguere le tipologie di caffè pone sul fondo delle tazze delle etichette identificative. Successivamente un assistente di laboratorio distribuisce ai degustatori una delle tazze, mentre il ricercatore osserva le loro espressioni facciali. a) L'assistente di laboratorio non dovrebbe guardare le tazze mentre vengono riempite.* b) L'assistente di laboratorio dovrebbe guardare le tazze mentre vengono riempite.
Causalità	Un ricercatore osserva che le regioni italiane con i parchi più grandi ospitano un numero maggiore di specie animali. a) Questi dati non dimostrano che l'aumento delle dimensioni dei parchi italiani porterà a un aumento del numero di specie animali.* b) Questi dati dimostrano che l'aumento delle dimensioni dei parchi italiani porterà a un aumento del numero di specie animali.
Variabili confondenti	Un ricercatore chiede a dei soggetti di risolvere un rompicapo. Alcuni di loro si trovano all'interno di una stanza fredda con una radio rumorosa, altri invece all'interno di una stanza calda e senza radio. I soggetti nella stanza calda senza radio risolvono il rompicapo più velocemente. a) Il ricercatore può affermare che sia la radio a causare una maggiore lentezza nella risoluzione del rompicapo.

	b) Il ricercatore non può affermare che sia la radio a causare una maggiore lentezza nella risoluzione del rompicapo.*
Validità di costrutto	Un ricercatore sta studiando le abilità matematiche generali di un campione di studenti, definite come un insieme di competenze che includono sia la risoluzione di problemi geometrici sia la capacità di eseguire calcoli complessi. a) Il ricercatore può misurare l'abilità generale in matematica utilizzando esclusivamente un test di geometria. b) Il ricercatore non può misurare l'abilità generale in matematica utilizzando esclusivamente un test di geometria.*
Gruppo di controllo	Due ricercatori testano una crema per l'acne su alcuni adolescenti con acne. Il ricercatore A vorrebbe dare la crema a tutti gli adolescenti presenti nello studio. Il ricercatore B vorrebbe darla solo a metà di loro, dando all'altra metà una crema neutra senza gli ingredienti per l'acne. a) Somministrare la crema a tutti gli adolescenti presenti nello studio è il modo migliore per verificare se la crema è efficace contro l'acne. b) Somministrare la crema solo a metà degli adolescenti, dando all'altra metà una crema neutra, è il modo migliore per verificare se la crema è efficace contro l'acne.*
Validità ecologica	Un ricercatore chiede a dei soggetti di giocare ad un gioco competitivo. L'obiettivo di ogni soggetto è guadagnare denaro comprando e vendendo gettoni. I partecipanti vengono pagati una tariffa fissa per partecipare all'esperimento. a) Il ricercatore può sostenere che il comportamento nell'esperimento rifletterà il comportamento di acquisto e vendita nella vita reale. b) Il ricercatore non può sostenere che il comportamento nell'esperimento rifletterà il comportamento di acquisto e vendita nella vita reale.*
Effetto della storia	Un campione di italiani selezionato casualmente viene intervistato sulla malattia A prima e dopo una campagna mediatica di 6 mesi sulla malattia. A metà della campagna mediatica, una celebrità italiana muore della malattia A. I dati dell'indagine indicano che la conoscenza della malattia A è maggiore dopo la campagna.

	a) Non è certo che la campagna mediatica abbia incrementato la conoscenza della malattia A.* b) È certo che la campagna mediatica abbia incrementato la conoscenza della malattia A.
Maturazione	Alcuni soggetti in un esperimento devono premere un pulsante ogni volta che un punto blu lampeggia sullo schermo del loro computer. Inizialmente, il compito è semplice; tuttavia, man mano che continuano a svolgerlo, i soggetti commettono sempre più errori. a) Possiamo stabilire che il punto blu lampeggi più velocemente man mano che il compito progredisce. b) Non possiamo stabilire che il punto blu lampeggi più velocemente man mano che il compito progredisce.*
Assegnazione casuale alle condizioni	Alcuni ricercatori vogliono verificare se un intervento psicoeducativo aiuta i bambini a migliorare le loro abitudini alimentari. I bambini che partecipano allo studio verranno divisi in un gruppo di intervento e in uno di controllo. a) I ricercatori dovrebbero assegnare casualmente metà dei bambini al gruppo di controllo e metà al gruppo di intervento.* b) I ricercatori dovrebbero assegnare i bambini che non fanno colazione al gruppo di intervento.
Affidabilità	Un ricercatore sviluppa un nuovo metodo per misurare la tensione superficiale dei liquidi. Questo metodo fornisce al ricercatore sempre lo stesso risultato. a) Rispetto al precedente, il nuovo metodo è anche necessariamente più preciso nel fornire una misura che corrisponda alla misura reale di tensione superficiale dei liquidi. b) Rispetto al precedente, il nuovo metodo non è necessariamente più preciso nel fornire una misura che corrisponde alla misura reale di tensione superficiale dei liquidi.*

The correct answers are marked with an asterisk.

Bibliography

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. (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), 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

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin, 52*(4), 281–302. https://doi.org/10.1037/h0040957

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

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

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

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

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), 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 prob-lem of internal consistency estimation. *British Journal of Psychology, 105*(3), 399–412. https://doi.org/10.1111/bjop.12046

Citizens' knowledge, perceptions, values and expectations of science – Report, Publications Office of the European Union, 2021. https://data.europa.eu/doi/10.2775/071577

European citizens' knowledge and attitudes towards science and technology – Eurobarometer report, Publications Office of the European Union, 2025. 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

Francis, L. J., & Crea, G. (2020). The psychometric properties of the Italian translation of the Astley-Francis Scale of Attitude toward Theistic Faith: A study across the age range 13- to 80-years. *Mental Health, Religion & Culture, 23*(3-4), 302–308. https://doi.org/10.1080/13674676.2020.1735324

Frederick, S. (2005). Cognitive reflection and decision making. *Journal of Economic perspectives*, *19*(4), 25-42. https://doi.org/10.1257/089533005775196732

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

Garnham, A., & Oakhill, J. (2013). *Mental models in cognitive science: Essays in honour of Phil Johnson-Laird*. Psychology Press. https://doi.org/10.1177/0013164406288165

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

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-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

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::AID-TEA6>3.0.CO;2-J

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.

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., 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

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

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

Meredith, W. (1993). Measurement invariance, factor analysis and factorial invariance. *Psychometrika*, *58*(4), 525-543. https://doi.org/10.1007/BF02294825

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/10.1080/07481756.1997.12068936

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13–103). Macmillan Publishing Co, Inc; American Council on Education.

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.

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

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., 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

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

Robba, M., Sorgente, A., & Iannello, P. (2024). In search of socially responsible investors: a Latent Profile Analysis. *Frontiers in Behavioral Economics*, *3*, 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 (Vol. 66, pp. 251-299). Academic Press. http://dx.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

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

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 Brian Everitt and David C. Howell. *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

Chapter 5 - Accumulating Validity Evidence for the Scientific Reasoning Scale in Italy: Integrating Classical Test Theory, Item Response Theory, and Latent Class Analysis[3]

Abstract: To contribute new validity evidence for the Italian Scientific Reasoning Scale (SRS) by integrating two complementary measurement frameworks – Classical Test Theory (CTT) and Item Response Theory (IRT) – with Latent Class Analysis (LCA). This triangulated approach allowed us to examine both scale-level properties and item-level quality, in line with the contemporary view of validity as the ongoing accumulation of evidence supporting score interpretations. The study re-analyzed data originally collected in a previous validation project of the Italian SRS, involving 897 adults recruited via quota sampling to approximate the Italian population. Within CTT, we tested the factor structure through confirmatory factor analysis and estimated composite reliability. In IRT, we compared Rasch/1PL, 2PL, and 3PL models, and assessed measurement precision using test information and EAP reliability. LCA was employed with an exploratory aim to evaluate item quality through five complementary indices (misclassified cases, classification error, Kullback–Leibler distance, item endorsement odds ratio, and class homogeneity). CTT and IRT analyses supported the unidimensional structure of the SRS and clarified its psychometric strengths and weaknesses: composite reliability was modest, and measurement precision was concentrated in the lower-to-mid range of the trait. Within CTT, Item 11 (response bias) showed insufficient factor loading on the latent factor, suggesting limited contribution to the construct. In IRT, we estimated a 2PL model, which highlighted variability in both discrimination and difficulty. Item 8 (maturation) emerged as highly discriminative, whereas Item 11 (response bias) showed low discrimination. In terms of difficulty, Item 9 (random assignment to conditions) was the easiest, capturing lower levels of scientific reasoning, while Item 11 (response bias) was the most difficult, targeting the upper end of the trait continuum. LCA confirmed that some items (e.g., Item 8, maturation) played a central role in classification accuracy, whereas others (e.g., Item 11, response bias) contributed less but nonetheless represent a theoretically important facet that warrants revision to strengthen construct coverage. The Italian SRS is a promising instrument, especially informative for distinguishing individuals in the lower-to-moderate range of scientific reasoning. Item 8 (maturation) consistently emerged as the strongest indicator across methods, whereas Item 11 (response bias) likely requires revision. Integrating CTT, IRT, and LCA strengthens the validity argument and offers practical guidance for scale refinement (e.g., adding more difficult items and assessing fairness via DIF).

[3] *This chapter is the result of a collaborative work conducted by Rossella Caliciuri, Angela Sorgente, and Margherita Lanz.*

Keywords: Scientific Reasoning Scale; Classical Test Theory; Item Response Theory; Latent Class Analysis; variable-centered approach; person-centered approach

1. Introduction

Scientific reasoning represents a cornerstone of scientific literacy and a critical competency for informed decision-making in modern societies (see previous chapter). It involves the ability to formulate and evaluate hypotheses, isolate and manipulate variables, and critically assess the strength of evidence and conclusions (Bao et al., 2009; Fischer et al., 2014). To capture these skills, Drummond and Fischhoff (2017) developed the Scientific Reasoning Scale (SRS), an 11-item instrument designed to assess multiple facets of the construct, including the understanding of blind/double-blind experiments, causality, confounding variables, construct validity, control group, ecological validity, history, maturation, random assignment to conditions, reliability, and response bias. Each item presents a brief scientific scenario and requires a true/false response.

In the previous chapter, we reported a study with an Italian sample that provided the first validity evidence for the SRS in this context. The study combined qualitative procedures (back-translation, expert reviews, and cognitive interviews to ensure linguistic and conceptual equivalence) with quantitative analyses based on Classical Test Theory (CTT). In line with the contemporary view of validity, multiple sources of evidence were collected using structural equation modelling (SEM): factorial structure through confirmatory factor analysis, reliability via omega coefficients, generalizability and known-group via measurement invariance between theoretically related groups (e.g., gender, age, education level, employment status, political orientation, and religious affiliation), and convergent and criterion through associations with theoretically related constructs (e.g., cognitive reflection, probabilistic reasoning, climate change awareness and beliefs, and paranormal health beliefs). These findings supported the intended score interpretations, indicating that the Italian SRS can be used to draw valid inferences about Italian adults' scientific reasoning. A key result of that study was the removal of the item addressing response bias, as it showed insufficient factor loading on the latent construct. While psychometrically justified, this decision also meant that the scale no longer represents this theoretically relevant facet, resulting in a more partial coverage of the construct. Building on this limitation, the present study seeks to deepen the evaluation of the excluded item by applying complementary methods such as Item Response Theory (IRT) and Latent Class Analysis (LCA), which can provide additional insights into its role and potential contribution to the construct despite its weaker Confirmatory Factor Analysis (CFA) performance.

According to the contemporary view of validity, validation is not a one-time achievement but an ongoing process of accumulating evidence across studies, populations, and contexts (AERA, APA & NCME, 2014; Hubley & Zumbo, 2011; Messick, 1989, 1995). This unified framework emphasises that there is only one form of validity, construct validity, and that multiple sources of evidence are needed to substantiate inferences from test scores. From this perspective, each new study contributes to a cumulative validity argument. Building on this view, the present work examines the Italian SRS through a broader methodological lens.

Over the past few decades, CTT and Item Response Theory (IRT) have frequently been presented as alternative, and sometimes competing, frameworks for psychometric evaluation (e.g., Kohli et al., 2015; see

also Fan, 1998, for an empirical comparison). CTT provides easily interpretable statistics, such as factor loadings, and internal consistency coefficients. However, its estimates are sample- and test-dependent, and unlike IRT, CTT does not allow for evaluating measurement precision across different levels of the latent trait (DeVellis & Thorpe, 2021). In IRT, by contrast, the probability of endorsing an item is modeled as a function of the latent trait. This approach yields item-level parameters (e.g., discrimination and difficulty) and provides information functions that describe measurement precision across different trait levels (Embretson & Reise, 2010). In recent years, however, the field has moved away from framing CTT and IRT as opposing paradigms, instead recognizing their complementary strengths (Bean & Bowen, 2021). Studies have increasingly employed both frameworks in parallel to obtain a more comprehensive assessment of item and scale quality (e.g., Bottesi & Spoto, 2025; Chen et al., 2025; Cui et al., 2025; Lu et al., 2025; Piumatti et al., 2021; Tong et al., 2025). This shift is also evident in research output: a Scopus search conducted in September 2025, using the keywords “item response theory” (or IRT) and “classical test theory” (or CTT) in titles, abstracts, and keywords, identified 1227 publications. As shown in Figure 1, the number of such studies has increased steadily, reflecting the growing recognition of their complementary value in psychometric research.

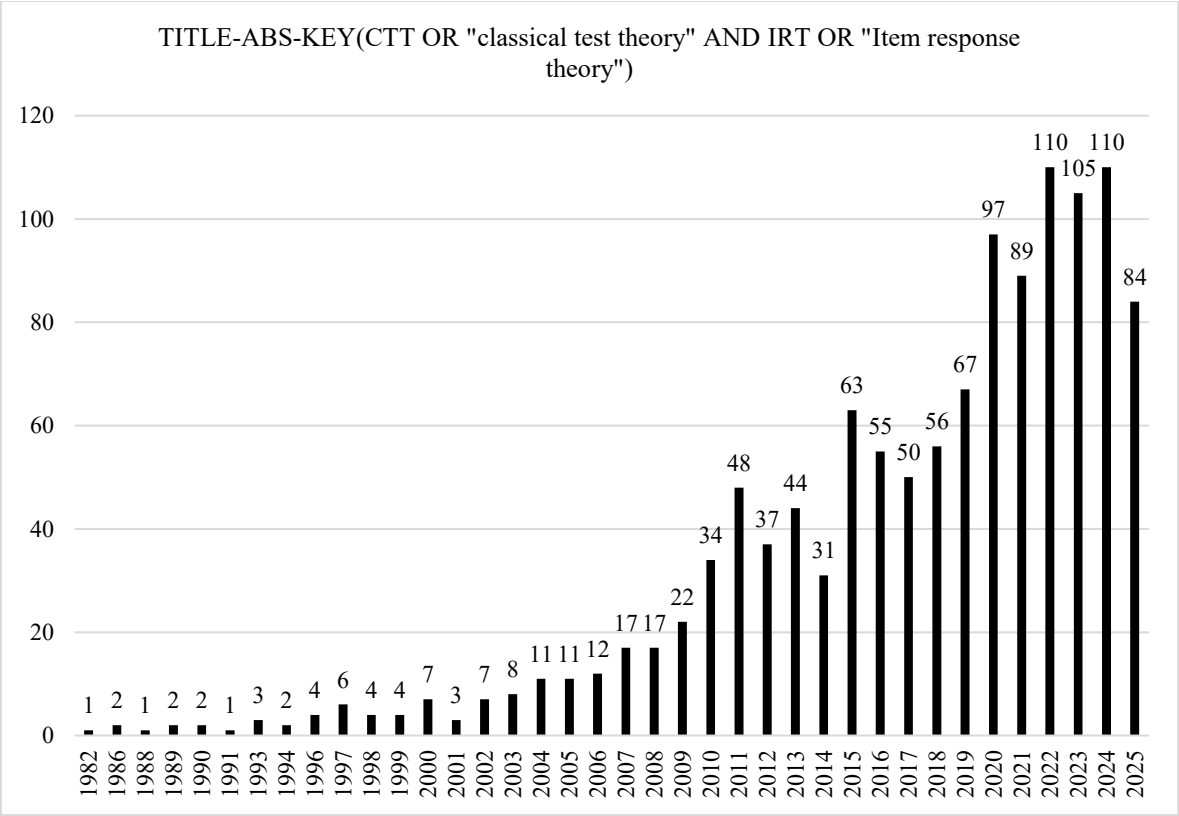

Fig. 1 *Number of publications using both Classical Test Theory and Item Response Theory approaches indexed in Scopus (titles, abstracts, and keywords). The bibliometric search, conducted in September 2025 yielded 1227 results*

Importantly, research has emphasized the advantages of integrating CTT and IRT, as each contributes unique information to the validation argument. According to Bean and Bowen (2021), CTT, particularly when

combined with CFA, offers accessible and well-established indicators of reliability and factorial structure, allowing researchers to test hypothesized measurement structures, evaluate model fit, and examine sources of error variance. By contrast, IRT provides fine-grained, model-based evidence of item performance and measurement precision, yielding parameters of discrimination and difficulty, item and test information functions, conditional standard errors, and person-specific reliability estimates. Bean and Bowen (2021) further demonstrate that while some information overlaps between the two measurement models, each also produces unique insights: CFA identifies how items relate to latent factors and whether alternative structural models fit the data, whereas IRT shows where along the latent continuum items provide the most information and highlights potential gaps in construct coverage. In sum, using CTT and IRT in combination allows for a more nuanced and complete understanding of how scales function across different populations, both at the scale and item level, thereby strengthening the overall validity argument.

1.1 Variable-centered and person-centered approaches

In addition to these established frameworks, we broaden the methodological scope by incorporating Latent Class Analysis (LCA) as an exploratory tool for item evaluation (Sorgente et al., 2025). Psychometric models can be broadly classified as variable-centered or person-centered, depending on their underlying logic and assumptions. Variable-centered frameworks, such as CTT and IRT, are theoretical paradigms designed to study the relationships among variables under the assumption that all individuals are drawn from a single population for which a single set of parameters can be estimated (Lanza & Cooper, 2016). Within these paradigms, specific data-analytic techniques (e.g., regression, confirmatory factor analysis) are employed to test hypotheses and evaluate measurement models, providing summary information that applies to the population as a whole (Howard & Hoffman, 2018). By contrast, person-centered approaches—such as Latent Class Analysis (LCA), Latent Profile Analysis (LPA), or cluster analysis—are based on the assumption that the sample may consist of multiple subpopulations, each characterized by distinct sets of parameters (Lanza & Cooper, 2016). In particular, LCA is a model-based clustering technique that classifies individuals into unobserved subgroups based on their response patterns, under the assumption that population heterogeneity can be explained by a finite number of discrete latent classes (Masyn, 2013). The aim of person-centered models is therefore to classify individuals into latent groups defined by similar response patterns, thereby emphasizing differences between groups of people rather than relationships among variables (Howard & Hoffman, 2018). These two perspectives are not mutually exclusive but can provide complementary insights. Previous research has often combined them—for example, using variable-centered analyses to examine the dimensionality of constructs and person-centered analyses to identify distinct subgroups (Morin et al., 2017; Meeusen et al., 2018; Lim & Yoon, 2025). However, such integration has rarely been pursued with the explicit aim of scale and item evaluation.

From this standpoint, the evaluation of item quality can also take a person-centered perspective. A high-quality item is one that meaningfully differentiates among classes, thereby increasing the likelihood of correctly classifying individuals into their latent class. In the context of item evaluation, this perspective offers an

alternative to traditional psychometric models: while CTT evaluates items based on factor loadings and reliability indices, and IRT emphasizes parameters of item discrimination and difficulty, LCA assesses item utility in terms of its ability to separate individuals into theoretically meaningful subpopulations. Although rarely applied for this purpose, some contributions have suggested that person-centered methods such as LCA can provide useful insights for item selection (Bartolucci et al., 2016; Cheng, 2009; Knack & Macready, 1985; van Buuren & Eggen, 2017). Given its limited use in this domain, the application of LCA should be regarded as exploratory, offering preliminary but potentially valuable insights that can complement more established psychometric approaches.

1.2 The present study

Within the contemporary framework of validity, understood as the ongoing accumulation of evidence to support test score interpretations, this study provides further validity evidence for the Italian SRS at the scale and item level. Building on the previous study, whose results led to the removal of the item addressing response bias due to its insufficient factor loading, we sought to further investigate this open issue while at the same time examining the functioning of the entire scale and all its items. To this end, we broadened the methodological scope by integrating CTT, IRT, and selected procedures from LCA, with the aim of triangulating findings and offering fine-grained insights into scale and item functioning. In doing so, we aim to clarify the role of the excluded facet, strengthen the overall validity argument, and advance the use of the scale in assessing scientific reasoning.

This study offers several novel contributions. It is the first to apply both CTT and IRT to the SRS, thereby extending previous work and providing a more comprehensive body of evidence in support of its use. By employing LCA for item evaluation, the study also introduces a less conventional yet promising perspective, situating the work at the intersection of person-centered and variable-centered approaches. Through this integrated framework, the study not only broadens the empirical foundation for the Italian SRS but also contributes methodologically by demonstrating the value of combining CTT, IRT, and LCA in scale validation. In line with the unified and ongoing view of validity (AERA et al., 2014; Messick, 1989), this work exemplifies how diverse methodological approaches can converge to strengthen the interpretation and practical use of psychological measures.

2. Methods

2.1 Participants

The present study draws on the same dataset used in the previous validation of the Italian SRS (see Chapter 4). Data were collected via Norstat, an online research platform where participants receive compensation for taking part in research studies. The sample was recruited through quota sampling, a non-probability technique designed to approximate population representativeness by dividing the population into strata (e.g., age, gender, and geographical region) and setting quotas proportional to national distributions. Participants provided

informed consent prior to data collection. The final sample comprised 897 adults (50.50% women; age range = 18–60; M = 41.48, SD = 12.46, median = 43). More detailed sociodemographic characteristics and recruitment procedures are reported in the previous chapter. The dataset is openly accessible on OSF (https://osf.io/6xrw7).

2.2 Procedure

For the present study, no new data were collected; existing data from the initial validation project of the Italian SRS were analysed with a different research objective. The project was preregistered on OSF in December 2024 and approved by the Ethical Committee of the Department of Psychology of Università Cattolica del Sacro Cuore, in accordance with the Declaration of Helsinki. The participants whose data were analysed in the present study had previously completed an online survey administered via Qualtrics between December 2024 and January 2025. Participation was voluntary, compensated according to the platform's guidelines, and contingent on informed consent. As part of the consent process, participants also agreed that their anonymized data could be made publicly available and reused for further research purposes.

2.3 Measures

The Scientific Reasoning Scale (SRS; Drummond & Fischhoff, 2017) was used to assess participants' scientific reasoning skills. For a detailed description of the instrument, its structure, and adaptation process, see the previous chapter.

2.4 Data analysis

2.4.1 Classical Test Theory (CTT)

Software and estimation. We ran CTT analyses (CFA) in Mplus (version 7.11; Muthén & Muthén, 1998–2017). We used the weighted least squares mean and variance adjusted estimator (WLSMV), which is appropriate for dichotomous indicators; missingness was handled with Mplus's default for categorical data under Missing At Random (MAR; pairwise present).

As described in the previous chapter, model fit was evaluated with chi-square, RMSEA, CFI, TLI, SRMR, and WRMR, following conventional guidelines for interpretation (e.g., Asparouhov & Muthén, 2018; Kline, 2023; Hu & Bentler, 1999). After establishing adequate model fit, reliability was assessed via composite reliability (ω), in line with current recommendations (Dunn et al., 2014). Further details on the model fit indices and their interpretation are reported in the previous chapter.

2.4.2 Item Response Theory

Software & estimation. Item response theory analyses were conducted in R (R Core Team, 2025) with the mirt package (Chalmers, 2012). Models were estimated via marginal maximum likelihood (MML) using the expectation–maximization algorithm (EM), which uses all available item responses without listwise deletion.

Missing data were negligible (0.00%–0.22%) and therefore not imputed. For the Orlando–Thissen S–χ^2 item-fit statistic, mirt applies row-wise listwise deletion, resulting in an effective sample size of N = 893.

Prior to estimating unidimensional IRT models, we examined the assumptions of unidimensionality and local dependence (LD), that is, residual covariation among item responses not accounted for by a single latent trait (Primi et al., 2015). Unidimensionality was evaluated by fitting a one-factor model and inspecting standardized factor loadings, with values ≥ .30 and statistical significance considered indicative of substantive loadings. LD was investigated using the χ^2 LD statistic (Chen & Thissen, 1997), with values ≥ 10 regarded as evidence of noteworthy local dependence.

To identify the most appropriate IRT model for our binary items, we considered three logistic models that differ in how many item parameters they estimate (Birnbaum, 1968; Lord, 1980). In the Rasch/1PL model, all items share the same discrimination (slope) and differ only in difficulty. In the 2PL model, each item has its own discrimination (*a*) and difficulty (*b*), allowing items to vary in how sharply they differentiate between respondents. The 3PL model adds a guessing (lower-asymptote) parameter (*c*), which represents a non-zero chance of endorsing the keyed response even at very low levels of the latent trait. We evaluated models using two complementary criteria. First, for nested pairs (Rasch vs. 2PL; 2PL vs. 3PL), we used the likelihood-ratio test based on the difference in −2 log-likelihoods (Δ−2LL), with a significant test supporting the more complex model. Second, we compared the Akaike Information Criterion (AIC; Akaike, 1974) and the Bayesian Information Criterion (BIC; Schwarz, 1978), with lower values indicating a better fit. For the 3PL, we also required proper convergence, well-defined standard errors, and a plausible estimate of the guessing parameter (*c*), as values that are excessively high are considered implausible in practice (Hambleton & Swaminathan, 1985; Embretson & Reise, 2000).

Global fit was evaluated using the M^2 statistic and the Root Mean Square Error of Approximation (RMSEA; Maydeu-Olivares & Joe, 2005, 2014). Adequate fit was defined by a nonsignificant M^2 ($p > .01$) and RMSEA ≤ .08. Item-level fit was tested with S–χ^2 statistics (Orlando & Thissen, 2000), with nonsignificant values ($p > .01$) indicating adequate item fit under the 2PL.

The 2PL logistic model (Birnbaum, 1968) was retained; this links the probability of a keyed response to the latent trait level (θ) via two item parameters: discrimination (*a*), indexing how sharply an item differentiates persons along θ, and difficulty/severity (*b*), locating the point on the θ-scale where the probability of endorsing the item is 50% (Nguyen et al., 2014). Discrimination magnitudes were interpreted using Baker & Kim's (2004) descriptive guidelines (low <.64; moderate .65–1.34; high 1.35–1.69; very high ≥1.70).

Measurement precision across θ was described with the test information function (TIF): information is inversely related to the squared standard error, so higher TIF implies smaller SE and greater precision at that θ level. Graphically, the TIF shows how well the construct is measured at different levels of the underlying construct continuum. Person scores were estimated using Expected a Posteriori (EAP) under the selected IRT model. To summarize average measurement precision, we report the EAP reliability, an IRT-based index that

compares the variability of EAP scores with their average posterior error variance; larger values indicate greater precision on average. Because test information varies across θ, this index can be modest when a test concentrates information in a narrow trait region and is lower elsewhere (Bock & Mislevy, 1982; Embretson & Reise, 2000).

2.4.3 Latent Class Analysis

Software and estimation. Latent Class Analysis was conducted in Mplus 7.11 (Muthén & Muthén, 1998–2017) using robust maximum likelihood estimation (MLR) with the Expectation–Maximization algorithm (EM). Analyses were based on the observed categorical indicators and missing data were handled with full information maximum likelihood (FIML).

Although latent class models are usually employed to identify subgroups of individuals, in the present study our main focus was on evaluating item quality. To apply item-quality indices in a meaningful way, however, it was first necessary to identify the latent class solution that adequately fit the data. Therefore, model selection and classification diagnostics were conducted as preliminary steps, followed by specific analyses of item quality.

To detect the number of latent groups (i.e., classes) that best captured the heterogeneity of scientific reasoning in the sample, we performed a Latent Class Analysis (LCA) including all 11 items of the SRS as observed indicators. Competing models were estimated sequentially, starting from a one-class solution and progressively increasing the number of classes until issues of under-identification or convergence emerged.

Model selection and evaluation of classification quality followed the steps and criteria proposed by Sorgente et al. (2019; 2025), considering both absolute and relative fit indices as well as standard diagnostics of classification accuracy (e.g., entropy, class proportions, average posterior probabilities, and odds of correct classification; see also Masyn, 2013).

After model selection, LCA was used to evaluate item quality by applying five complementary procedures, grouped into two overarching approaches. The first approach focused on exclusion effects, based on the idea that high-quality items should worsen the solution if excluded from the item pool. Two indices were used for this purpose. The proportion of misclassified cases (Bartolucci et al., 2016) quantifies the number of individuals whose class membership changes when an item is removed. To compute this index, the LCA was re-estimated while excluding one item at a time and the resulting class assignments were compared with those from the full 11-item model; the larger the number of reassignments, the more important the item is for preserving stable classification. The second indicator was the classification error index (Knack & Macready, 1985), which represents the complement of the average posterior probability of correct classification. It is obtained by weighting the average posterior probability of correct assignment by the relative size of each class and subtracting this value from one. The index was calculated for the full model and for each reduced model excluding one item. The difference between these values was then considered: a positive difference indicates

that removing the item increases classification error – and the larger this increase, the more the solution deteriorates – whereas a negative difference suggests a slight improvement in classification accuracy.

The second approach focused on class differentiation, according to which high-quality items are those characterized by substantial differences in endorsement probabilities across latent classes. Different indices were employed. The Kullback–Leibler Distance (Cheng, 2009; van Buuren & Eggen, 2017) was computed by comparing, for each item, the conditional response probability distributions across classes. Higher values indicate that the response distribution for that item diverges more strongly between classes, reflecting a greater contribution to class separation. The item endorsement odds ratio, proposed by Masyn (2013), was obtained by taking the ratio of the probability of endorsing an item in one class relative to another; large odds ratios indicate that endorsement probabilities differ substantially across classes, making the item particularly informative for distinguishing latent profiles. However, Masyn also stressed that class separation is not sufficient on its own: items should additionally show class homogeneity, that is, endorsement probabilities within each class reaching extreme values (i.e., greater than .70 or lower than .30). While homogeneity indicates that members of a given class respond consistently, an item can only be considered high quality when both conditions are satisfied: it separates classes (odds ratio) and shows homogeneity within classes.

3. Results

3.1 Classical Test Theory

As described in the previous chapter, confirmatory factor analysis supported the unidimensionality of the SRS but highlighted weaknesses in Item 11 (response bias), whose factor loading fell below the recommended threshold. For this reason, the best-fitting solution within the CTT framework involved removing this item, resulting in a 10-item scale. Composite reliability for the revised scale was $\omega = .612$.

3.2 Item Response Theory

Results showed that a single factor model adequately represents the structure of SRS given that none of the LD statistics were greater than 10. All factor loadings were significant ($p < .001$) and ranged from .392 to .656. Only item 11 (response bias) fell below the conventional .30 loading threshold (.185) and should be interpreted with caution; all other items exceeded .30.

We estimated 1PL (Rasch), 2PL, and 3PL models. The Rasch model showed acceptable fit [M_2 (54) = 152.10, $p < .001$; RMSEA = .045 (.037 - .053)], as did as the 2PL [M_2 (44) = 74.93, $p = .002$; RMSEA = .028 (.017 - .039)]. Attempts to fit the 3PL did not yield a usable solution (non-convergence or implausible guessing parameter, $c \approx .50$). The 2PL significantly improved over the Rasch model [$\Delta-2LL(10) = 75.585$, $p < .001$], and it had lower BIC (11267.67 vs. 11275.28) and AIC (11162.12 vs. 11217.70). We therefore retained the 2PL as the final model for the SRS.

Under the 2PL model, all items showed good fit (Table 1).

Table 1 *Item fit in the two-parameter logistic (2PL) model*[a]

item	S-χ^2	df	p (S-χ^2)
1	5.289	7	0.625
2	9.499	6	0.147
3	7.594	7	0.370
4	3.970	7	0.783
5	1.729	7	0.973
6	3.339	7	0.852
7	7.157	7	0.413
8	5.433	7	0.607
9	9.095	7	0.246
10	3.670	7	0.817
11	5.762	7	0.568

[a]*Note.* S-χ^2 = Orlando–Thissen item-fit chi-square statistic; df = degrees of freedom; p (S-χ^2) = p-value associated with the item-fit test, where nonsignificant values (p > .01) indicate adequate item fit under the 2PL model.

Table 2 reports the item parameters estimated under the 2PL model. Discrimination (*a*) ranged from 0.32 (item 11) to 1.48 (item 8) logit. Following Baker & Kim (2004), items 11 and 5 shows low discrimination (< .64); items 1, 2, 3, 4, 6, 7, 9, and 10 are moderate (.65–1.34); and item 8 is high (1.35–1.69). Difficulty (*b*) values were all below 0 on the θ metric (mean = 0, SD = 1), spanning −1.85 to −.19. Thus, the scale primarily targets low to medium levels of the latent trait. Item 9 was the easiest (b = −1.85), whereas item 11 was the hardest (b = −.19).

Table 2 *Item parameters under the two-parameter logistic (2PL) model*[b]

item	a	B
1	0.73	-1.26
2	0.88	-0.44
3	1.33	-1.09
4	1.25	-1.62
5	0.59	-1.51
6	1.29	-0.99
7	0.78	-1.06
8	1.48	-1.04
9	0.97	-1.85
10	0.89	-1.02
11	0.32	-0.19

[b]*Note.* a = item discrimination parameter, indexing how sharply an item differentiates respondents along the latent trait (θ); higher values indicate greater discrimination. b = item difficulty (location) parameter, representing the level of θ at which the probability of a correct response is 50%; negative values indicate items are easier and located below the mean of the latent trait distribution ($\theta = 0$).

Item parameters were visualized through Item Characteristic Curves (ICCs), which depict the logistic function relating the probability of a correct response to the level of the latent trait (θ). All items were located in the negative range of the trait, with discrimination reflected in the slope of the curve (steeper slopes indicate higher discrimination). The full set of ICCs is available in the online supplementary materials hosted on OSF (https://osf.io/6xrw7).

The Test Information Function (TIF; Figure 2) peaks at $\theta \approx -1.13$ (TIF ≈ 2.73, corresponding SE ≈ 0.60), with comparatively higher information across $\theta \in [-1.6, -0.5]$. This indicates that the SRS provides its most precise measurement for individuals scoring in the low-to-moderate range of the latent trait, while measurement precision decreases around the mean ($\theta \approx 0$) and at the extremes. The overall EAP reliability was ≈ 0.37, reflecting a test that is reasonably accurate near its information peak but substantially less precise outside this range.

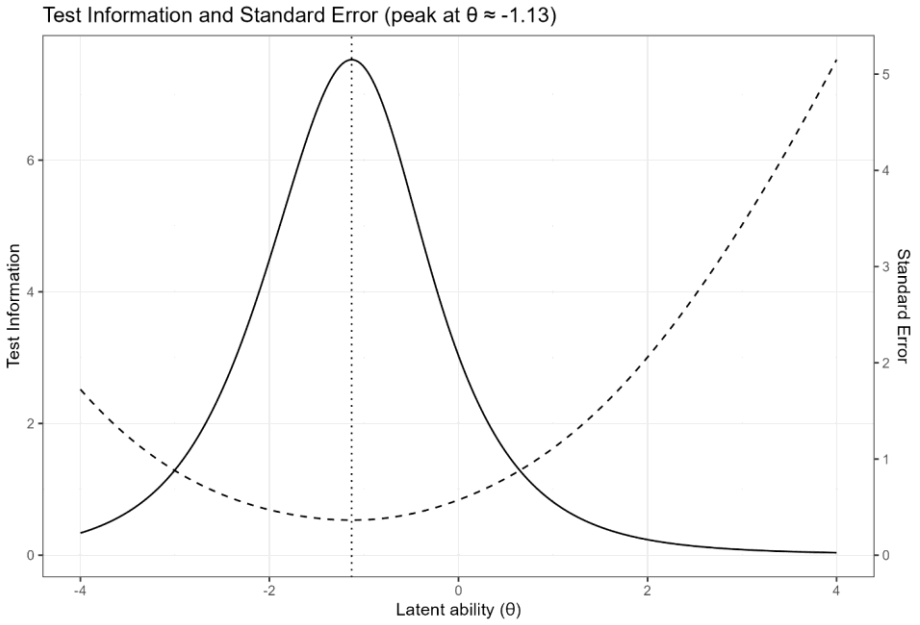

Fig. 2 *Test Information Function (TIF) and Standard Error across the latent trait*

Note. The solid line represents the Test Information Function (TIF), while the dashed line represents the Standard Error (SE). The dotted vertical line marks the location of the TIF peak ($\theta \approx -1.13$). Solid line: Test Information Function; dashed line: Standard Error (right axis). Precision is highest where information peaks ($\theta \approx -1.13$); uncertainty increases as information declines away from that region.

To gauge the impact of the very low-discrimination item, we re-estimated a 2PL model excluding Item 11 (response bias). Results were essentially unchanged relative to the full 2PL: the TIF peak was virtually identical

(full: $\theta \approx -1.1$, TIF ≈ 2.73; model without item 11: $\theta \approx -1.2$, TIF ≈ 2.72). Person scores (θ̂_EAP) were almost perfectly aligned (r = .997) with a mean absolute deviation (MAD, i.e., the average absolute difference between person estimates across models) of 0.062. EAP reliability decreased only slightly, from 0.369 to 0.358. Taken together, although Item 11 shows a weak loading (.185) and low discrimination (.32), its removal has negligible impact on the scale, leaving both the metric and information profile essentially unchanged around $\theta \approx -1$. The choice may therefore be guided by content and theoretical considerations; accordingly, within the IRT framework, we decided to retain Item 11 to preserve its substantive relevance.

3.3 Latent Class Analysis

We compared seven candidate models, and based on fit indices (Table 3) the 2-class solution emerged as the best-fitting model. This solution, which also showed adequate classification quality (Table 4), distinguished between respondents who were able to answer correctly only a subset of items (with average probabilities of correct responses around 50–60%) and those who answered correctly to almost all items (with probabilities often ranging from 70–95%). The 2-class solution was therefore retained, and the five complementary procedures were applied to evaluate item quality.

Table 3 *Absolute and Relative Model Fit Indices for Six Latent Class Models*[c]

Model	LL	χ^2_{LRT}	stdres	AIC	CAIC	AWE	BIC	SABIC	SIC	BF	cmP	VLMR-LRT test	LMR-LRT test	BLRT test
1-class	-5816.65	p= 1.00	23.01%	11655,30	11719,08	11815,85	11708,08	11673,14	-5854,04	0.00	0.00			
2-class	-5556.47	p= 1.00	11.45%	11158,94	11292,29	**11494,65**	**11269,29**	**11196,25**	-5634,65	**3569325494,78**	**1.00**	**p < .001**	**p < .001**	**p < .001**
3-class	-5537.68	p= 1.00	11.67%	11145,36	11348,29	11656,21	11313,29	11202,13	-5656,64	**228692119104,97**	1.00	p = .3938	p = .3992	p = .0128
4-class	-5523.05	p= 1.00	9.91%	11140,09	11412,60	11826,10	11365,60	11216,33	-5682,80	**3125584675726,02**	0.00	p =.4173	p = .4211	p = .1818
5-class	-5511.03	p= 1.00	9.03%	11140,06	11482,14	12001,22	11423,14	11235,76	-5711,57	**2976122747503,91**	0.00	p = .5170	p = .5194	p = .2667
6-class	-5498.96	p= 1.00	9.91%	**11139,93**	11551,58	12176,24	11480,58	11255,10	-5740,29	**13034782506787,40**	0.00	p = .3445	p = .3468	p = 1.000
7-class	**-5488.37**	p= 1.00	11.23%	11142,75	**11059,75**	12354,21	11540,98	11277,39	-5770,49	/	0.00	p = .7234	p = .7247	p = .6667

[c]*Note*. LL = loglikelihood; χ^2_{LRT} = likelihood ratio chi square goodness of fit; stdres = standardized residuals; AIC = Akaike information criterion; CAIC = Consistent AIC; AWE = Approximate Weight of Evidence criterion; BIC = Bayesian information criterion; SABIC = sample-size adjusted BIC; AIC = Akaike information criterion; BIC = Bayesian information criterion; BF = Bayesian factor; cmP = approximate correct model probability; VLMR-LRT = Vuong-Lo-Mendell-Rubin likelihood ratio test; LMR-LRT = Lo–Mendell–Rubin likelihood ratio test; BLRT = bootstrapped likelihood ratio test. Fit indices corresponding to the best-fitting model(s) are shown in bold.

Table 4 *Classification Diagnostics for the Two-Class Model*[d]

Entropy (E)	Class (N)	CP	mcaP	AvePP	OCC
0.629	Class 1 (410)	.464 (.387 - .539)	0.464	0.889	9.25
	Class 2 (486)	.536 (.461 - .613)	0.536	0.894	7.30

[d]*Note*. E = Entropy; CP = class proportion; mcaP = modal class assignment proportion; avePP = average posterior probability; OCC = odds of correct classification.

Procedures based on exclusion effects. The proportion of misclassified cases (Bartolucci et al., 2016) showed that eliminating Item 8 (maturation) had the strongest impact on classification accuracy, with 73 individuals reassigned to different classes. In contrast, removing Item 11 (response bias) led to only 16 misclassifications, indicating its marginal role in preserving the latent structure. The classification error indices (Knack & Macready, 1985) revealed that overall misclassification error slightly decreased when Item 1 (blind/double-blind) was excluded, whereas it increased when Item 8 (maturation) was removed, confirming the central role of this item in maintaining model stability.

Procedures based on class differentiation. The Kullback–Leibler divergence (Cheng, 2009; van Buuren & Eggen, 2017) indicated that Item 1 (blind/double-blind) contributed the most to differentiating latent response distributions (KL = .064), while Item 11 (response bias) contributed the least (KL = .003). Consistently, the odds ratios of endorsement probabilities (Masyn, 2013) showed that Item 11 (response bias) had the weakest ability to separate classes, whereas all other items produced comparable odds ratios. The class homogeneity index (Masyn, 2013) showed that Items 3, 4, 6, 8, 9 (confounding variables, construct validity, ecological validity, maturation, random assignment to conditions) were characterized by sufficiently extreme endorsement probabilities within classes, marking them as distinctive indicators of class-specific response patterns.

Across methods (Table 5), Item 8 (maturation) consistently emerged as the most informative and discriminative, while Item 11 (response bias) was the weakest, providing little contribution to either classification accuracy or class separation. Other items, such as 1, 3, 4, 6, and 9 (blind/double blind, confounding variables, construct validity, ecological validity, random assignment to condition), also showed solid performance across multiple criteria. These convergences suggest that the scale contains a core subset of items with strong discriminative properties, while one item (i.e., Item 11) in particular appears redundant or less informative.

Table 5 *Item quality indices from Latent Class Analysis*[e]

Item	Facet	Misclassified cases	Classification error	KL Distance	Item Endorsment Odds Ratio	Class Homogeneity
1	Blind/double blind	29	-0.000215	0.063501	1	0
2	Causality	54	0.006491	0.053813	1	0
3	Confounding variables	60	0.013964	0.038125	1	1
4	Construct validity	55	0.007007	0.034390	1	1
5	Control group	33	0.002853	0.024623	1	0
6	Ecological validity	66	0.011980	0.024122	1	1
7	History	42	0.006712	0.023509	1	0
8	Maturation	73	0.019684	0.020929	1	1
9	Random assignment to condition	38	0.008444	0.015103	1	1
10	Reliability	48	0.004471	0.014769	1	0
11	Response bias	16	0.000257	0.003259	0	0

[e]*Note.* Misclassified cases = number of individuals reassigned to a different class when the item is excluded; higher values identify high-quality items because they indicate stronger impact on classification accuracy (Bartolucci et al., 2016). Classification Error = change in overall classification error relative to the full model; increases identify high-quality items, as their exclusion worsens classification accuracy and increased misclassification (Knack & Macready, 1985). KL Distance = Kullback–Leibler divergence quantifying the information loss when excluding the item; higher values identify high-quality items, as they contribute more to class separation (Cheng, 2009; van Buuren & Eggen, 2017). Item Endorsement Odds Ratio = relative likelihood of endorsing the item across classes; larger values identify high-quality items, because they show stronger differences between classes (Masyn, 2013). Class Homogeneity = whether endorsement probabilities within classes were sufficiently extreme (> .70 or < .30), coded as 1 = yes, 0 = no; values of 1 identify high-quality items, as they show coherent response patterns within each class (Masyn, 2013).

4. Discussion

The present study aimed to provide new validity evidence for the Italian version of the Scientific Reasoning Scale (SRS) by integrating two complementary measurement frameworks—Classical Test Theory (CTT) and Item Response Theory (IRT)—with Latent Class Analysis (LCA), a person-centered data-analytic technique

that was here applied in an exploratory way to evaluate item quality. While CTT and IRT are established models of measurement and were used to gain a comprehensive understanding of the scale and its items, LCA offered a novel and still rarely applied perspective for assessing whether specific items meaningfully contribute to the differentiation of response patterns. Situating our work within the contemporary framework of validity (AERA, APA, & NCME, 2014; Messick, 1989, 1995), this study exemplifies how accumulating evidence from multiple approaches can converge to strengthen the validity argument for a psychological measure.

4.1 Main findings at the scale and item level

At the scale level, both CTT and IRT analyses confirmed the unidimensionality of the SRS and supported its overall structural coherence, although internal consistency was modest and the measurement precision was concentrated in the lower to mid ranges of the latent trait. The Test Information Function peaked around $\theta \approx -1$, indicating that the scale is most accurate in distinguishing individuals with low to moderate levels of scientific reasoning, whereas sensitivity decreases at higher ability levels.

At the item level, converging evidence from CTT, IRT, and LCA highlighted a subset of items with strong psychometric properties, alongside others that appear weaker. In particular, item 8 (maturation) consistently emerged as the strongest indicator across all approaches: it showed the highest factor loading in CTT, high discrimination in IRT, and the largest impact on classification indices in LCA, confirming its centrality for the construct. Item 9 (random assignment to conditions) emerged as the easiest item according to IRT ($b = -1.85$), making it primarily informative at the lower end of the trait. In the previous CTT-based study realized with the same data used in the current study, for strong invariance in the gender comparison this item ("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") had to be freed: females were more likely than males to answer incorrectly despite equivalent latent trait levels. This combination of findings indicates that Item 9 should be carefully monitored with future DIF analyses in IRT, as it appears to be a theoretically important but psychometrically sensitive indicator. Item 11 (Response bias) showed consistently weak performance: a very low factor loading in CFA, minimal discrimination in IRT, and negligible contribution to LCA indices. Despite this, the item should not be removed solely on psychometric grounds; a content-informed revision is preferable to maintain construct coverage, since it represents a theoretically central facet of scientific reasoning. Instead, reformulation is needed to improve its discriminative capacity while retaining coverage of this dimension. Finally, other items such as blind/double blind (Item 1), confounding variables (Item 3), construct validity (Item 4), and ecological validity (Item 6) performed well across methods. By contrast, control group (Item 5), history (Item 7), causality (Item 2), and reliability (Item 10) showed somewhat less consistent evidence across the different approaches: their factor loadings and discrimination parameters were generally acceptable but not as strong as those of the best-performing items, and their contribution to class differentiation in LCA was more limited. These findings do not undermine their theoretical relevance, but suggest that closer monitoring of their functioning may be warranted in future studies.

4.2 Implications for the SRS

The joint use of CTT, IRT, and LCA should be understood not only as a methodological comparison but also as a way of refining the theoretical interpretation of scientific reasoning as measured by the SRS. Across models, the scale appears to capture the construct at a relatively basic level: it is informative in distinguishing lower and mid levels of competence, but less so for more advanced expressions of scientific reasoning. This suggests that the current item set measures important facets of scientific reasoning, but only at a basic level of difficulty, without adequately covering more demanding manifestations of the construct.

At the item level, some items consistently perform well (e.g., maturation), while others show weaknesses across methods (e.g., response bias) and require revision. More broadly, our findings suggest that each of the eleven facets of scientific reasoning should be represented by more than a single item. Currently, the SRS relies on one item per facet, which restricts its ability to capture variability within each domain. Future versions should therefore include multiple items per facet, designed to cover different levels of difficulty, so that the scale can more effectively represent the full spectrum of scientific reasoning.

From this perspective, the exploratory use of LCA adds value by evaluating item quality in terms of their contribution to classification accuracy. Although preliminary, these results complement the information obtained from CTT and IRT and highlight how person-centered indices can support item refinement.

Taken together, these findings suggest that the Italian SRS is a promising instrument for assessing scientific reasoning, particularly useful for detecting differences among individuals in the lower to middle ranges of the ability spectrum. This makes it a valuable tool for educational research and interventions aimed at identifying areas of weakness and fostering the development of reasoning skills in the general population. At the same time, the limited precision at higher ability levels suggests the need to enrich the scale with more challenging items ($b > 0$) and possibly to develop additional items within the existing facets of the construct, so as to broaden coverage and balance measurement precision across the trait continuum.

Future refinements of the SRS should consider two practical directions: developing additional items with higher difficulty parameters ($b > 0$) to extend measurement precision into the upper range of the trait, and creating multiple items for each theoretical facet rather than relying on a single indicator, so as to capture intra-facet variability and strengthen construct coverage. At the same time, it is important to recognize that validity evidence is always sample- and context-specific; this study adds new evidence for Italian adults, thereby contributing to the cumulative validation process of the SRS across populations.

4.3 Broader implications for scale development

Beyond the case of the SRS, our findings speak to the general process of test construction and validation. As emphasized by Messick (1989, 1995) and reiterated in the Standards for Educational and Psychological Testing (AERA, APA, & NCME, 2014), validation is not a purely statistical task but an integrated effort requiring coherence between psychometric evidence and theoretical representation of the construct. This position has

been further elaborated by Hubley and Zumbo (2011), who describe validation as a cumulative and ongoing process in which multiple sources of evidence are progressively accumulated, each contributing information that is specific to the context and sample in which it is obtained. Their framework highlights that statistical adequacy alone is insufficient: an item can display good psychometric properties while failing to represent an essential dimension of the construct, or conversely, a theoretically central item may show weak statistical performance but still need to be retained and refined in order to preserve conceptual coverage.

Recent studies echo this view. For instance, Tong et al. (2025) and Bottesi & Spoto (2025) show how combining CTT and IRT allows more informed decisions about the theoretical adequacy of instruments. Similarly, Cui et al. (2025) emphasize that statistical indicators must be interpreted alongside content validity, so that each retained item contributes both numerically and conceptually to construct representation.

Taken together, these contributions underscore a key principle: test development must balance psychometric rigor with theoretical coverage. A well-functioning scale is not one that merely optimizes fit indices or information functions, but one in which psychometric properties and conceptual relevance are jointly considered. In the case of the SRS, this means that future refinements should not be guided only by the statistical performance of items, but also by their ability to represent all theoretically important facets of scientific reasoning at different levels of difficulty. Only by holding together the psychometric and theoretical dimensions can the scale achieve both statistical robustness and conceptual completeness.

A further implication of our work relates to the broader landscape of measurement in psychology. Recent commentaries have pointed out that the field suffers from what has been termed the “toothbrush problem”: the tendency for researchers to constantly develop new instruments rather than reusing and refining existing ones (Elson et al., 2023). This proliferation of measures fragments the literature, complicates comparisons across studies, and ultimately undermines the cumulative accumulation of evidence. By contrast, working to strengthen the validity of established scales promotes continuity and comparability, thereby supporting cumulative science. Our effort to refine the Italian SRS is therefore justified not only by the need for cultural adaptation but also by the importance of consolidating construct validity in an existing instrument, rather than creating yet another competing measure. Strengthening well-established tools enhances comparability across studies and fosters cumulative progress.

4.4 Methodological contributions

This study contributes methodologically in two main ways. First, it joins a growing body of research in which both CTT and IRT are used in parallel to evaluate scale properties and collect complementary sources of validity evidence (e.g., Bottesi & Spoto, 2025; Chen et al., 2025; Cui et al., 2025; Lu et al., 2025; Piumatti et al., 2021; Tong et al., 2025). The increasing prevalence of this integrated approach reflects the recognition that the two frameworks provide distinct but converging information: CTT offers accessible and well-established indicators of structure and reliability, while IRT adds precision about item functioning and measurement error across the trait continuum.

Second, we introduced LCA as an exploratory tool for item evaluation. Previous studies (e.g., Bartolucci et al., 2016) have already applied LCA-based methods for item selection, typically relying on one or a few indices. To our knowledge, however, this is the first study to integrate and systematically apply multiple indices within the same framework, thereby providing a more comprehensive picture of item quality. While preliminary, this approach demonstrates the potential of person-centered methods to complement traditional measurement models by offering insights into which items truly drive meaningful distinctions among respondents.

4.5 Limitations and future directions

The main limitation of this study is that the data come from the same collection used in the previous validation, which constrains the generalizability of findings. Replication with independent samples is therefore needed to confirm the robustness of the results. In addition to this limitation, several directions for future research emerge. One concerns the comparison of different approaches to invariance testing, namely measurement invariance analyses conducted with multi-group CFA (within the CTT framework) and Differential Item Functioning (DIF) detection based on IRT models. Such a comparison would clarify how these methods, though both designed to assess invariance, may converge or diverge in evaluating the stability and fairness of the scale across groups. Another promising avenue relates to the use of LCA-based indices for item evaluation. In the present study, these were implemented under the assumption that the initial 11-item version of the scale represented the maximally informative solution, with comparisons made between the full model and models omitting one item at a time. This assumption may limit the reliability of findings, as the quality indices depend on the adequacy of the starting solution. Future research should explore whether LCA-based indices yield consistent conclusions when applied to different starting solutions (e.g., alternative item pools, revised versions of the scale or comparisons with randomly simulated values), rather than assuming the full version to be optimal. Furthermore, future studies are needed to determine whether such procedures are better suited for the early stages of scale construction and validation (as we have done here to obtain insights into item quality) or for the refinement of long item pools where they may help in reducing redundancy and guiding the creation of short forms.

5. Conclusion

The present study strengthens the validity argument for the Italian Scientific Reasoning Scale by showing that it can capture meaningful differences in scientific reasoning among adults, particularly in the lower-to-moderate range of ability. While some items emerged as stronger indicators than others, the overall findings confirm that the scale represents a useful tool for studying an ability that is central to scientific literacy and informed decision-making.

From a methodological standpoint, the study illustrates the value of combining multiple psychometric perspectives. Classical Test Theory, Item Response Theory, and Latent Class Analysis each contributed complementary insights that would not have been visible through a single framework alone. This triangulated approach highlights how test evaluation can benefit from the integration of variable-centered and person-

centered methods, thereby enriching both the theoretical interpretation of constructs and the practical guidance for scale refinement.

These findings point to the importance of continuously refining existing instruments rather than multiplying new ones, and of grounding psychometric evaluation in both statistical evidence and theoretical coherence. In this way, measurement tools such as the SRS can evolve into increasingly robust resources for advancing cumulative knowledge in psychology and education.

Bibliography

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control, 19*(6), 716–723. https://doi.org/10.1109/TAC.1974.1100705

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

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

Baker, F. B., & Kim, S.-H. (2004). *Item response theory: Parameter estimation techniques* (2nd ed.). New York, NY: Marcel Dekker.

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

Bartolucci, F., Montanari, G. E., & Pandolfi, S. (2016). Item selection by latent class-based methods: an application to nursing home evaluation. *Advances in Data Analysis and Classification, 10*(2), 245-262. https://doi.org/10.1007/s11634-016-0232-3

Bean, G. J., & Bowen, N. K. (2021). Item Response Theory and Confirmatory Factor Analysis: Complementary Approaches for Scale Development. *Journal of Evidence-Based Social Work*, *18*(6), 597–618. https://doi.org/10.1080/26408066.2021.1906813

Birnbaum, A. (1968). *Some latent trait models and their use in inferring an examinee's ability.* In F. M. Lord & M. R. Novick (Eds.), *Statistical theories of mental test scores* (pp. 397-479). Addison-Wesley.

Bock, R. D., & Mislevy, R. J. (1982). Adaptive EAP estimation of ability in a microcomputer environment. *Applied Psychological Measurement, 6*(4), 431–444. https://doi.org/10.1177/014662168200600405

Bottesi, G., & Spoto, A. (2025). Should we worry about how we measure worry? Insights from an updated version of the Italian Penn State Worry Questionnaire. *International Journal of Clinical and Health Psychology*, *25*(2), 100579. https://doi.org/10.1016/j.ijchp.2025.100579

Chalmers, R. P. (2012). mirt: A multidimensional item response theory package for the R environment. *Journal of Statistical Software, 48*(6), 1–29. https://doi.org/10.18637/jss.v048.i06

Chen, P., Li, Y., Xiong, M., & Zhang, Y. (2025). Psychometric evaluation of the behavioral inhibition/activation system scales in older adults in Mainland China: A classical test theory and item response theory approach. *Acta Psychologica*, *258*, 105280. https://doi.org/10.1016/j.actpsy.2025.105280

Chen, W. H., & Thissen, D. (1997). Local dependence indexes for item pairs using item response theory. *Journal of Educational and Behavioral Statistics*, *22*(3), 265-289. https://doi.org/10.3102/10769986022003265

Cheng, Y. (2009). When Cognitive Diagnosis Meets Computerized Adaptive Testing: CD-CAT. *Psychometrika, 74*(4), 619–632. https://doi.org/ 10.1007/s11336-009-9123-2

Cui, J., Wang, J., Yue, A., Cao, J., Zhang, Z., & Shi, B. (2025). Psychometric evaluation of the Chinese version of the Patient Self-Advocacy Scale using classical test theory and item response theory. *Scientific Reports*, *15*(1), 6871. https://doi.org/10.1038/s41598-025-91129-2

DeVellis, R. F., & Thorpe, C. T. (2021). *Scale development: Theory and applications*. Sage publications. https://doi.org/10.1111/peps.12499

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 prob-lem of internal consistency estimation. *British Journal of Psychology, 105*(3), 399–412. https://doi.org/10.1111/bjop.12046

Elson, M., Hussey, I., Alsalti, T., & Arslan, R. C. (2023). Psychological measures aren't toothbrushes. *Communications Psychology*, *1*(1), 25. https://doi.org/10.1038/s44271-023-00026-9

Embretson, S.E., & Reise, S.P. (2000). *Item Response Theory for Psychologists* (1st ed.). Psychology Press. https://doi.org/10.4324/9781410605269

Fan, X. (1998). Item response theory and classical test theory: An empirical comparison of their item/person statistics. *Educational and psychological measurement*, *58*(3), 357-381. https://doi.org/10.1177/0013164498058003001

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

Hambleton, R. K., & Swaminathan, H. (1985). *Item response theory: Principles and applications.* Dordrecht: Springer.

Howard, M. C., & Hoffman, M. E. (2018). Variable-centered, person-centered, and person-specific approaches: Where theory meets the method. *Organizational Research Methods, 21*(4), 846–876. https://doi.org/10.1177/1094428117744021

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

Kline, R. B. (2023). *Principles and practice of structural equation modeling*. Guilford publications.

Knack, R. E., & Macready, G. B. (1985). LC-PROG: optimal item selection in a latent class classification paradigm. *Applied Psychological measurement,* 9(4), 432-432. https://doi.org/10.1177/014662168500900413

Kohli, N., Koran, J., & Henn, L. (2015). Relationships among classical test theory and item response theory frameworks via factor analytic models. *Educational and psychological measurement*, *75*(3), 389-405. https://doi.org/10.1177/0013164414559071

Lanza, S. T., & Cooper, B. R. (2016). Latent class analysis for developmental research. *Child Development Perspectives, 10*(1), 59–64. https://doi.org/10.1111/cdep.12163

Lim, A. Y., & Yoon, W. (2025). Integrating variable-centered and person-centered approaches for personality and nicotine use. *Scientific Reports, 15*, 5433. https://doi.org/10.1038/s41598-025-90042-y

Lo Y., Mendell N. R., Rubin D. B. (2001). Testing the number of components in a normal mixture. *Biometrika*, 88, 767–778. https://doi.org/10.1093/biomet/88.3.767

Lord, F. M. (1980). *Applications of Item Response Theory to Practical Testing Problems*. Hillsdale, NJ: Lawrence Erlbaum.

Lu, H., Li, X., & Li, K. (2025). Adaptation and Validation of the Scale for Chinese Preschool Teachers' Self-Efficacy (SCPTSE): Based on Classical Test Theory and Item Response Theory. *Behavioral Sciences*, *15*(6), 741. https://doi.org/10.3390/bs15060741

Maydeu-Olivares, A., & Joe, H. (2005). Limited- and Full-Information Estimation and Goodness-of-Fit Testing in 2^n Contingency Tables: A Unified Framework. *Journal of the American Statistical Association*, *100*(471), 1009–1020. http://www.jstor.org/stable/27590631

Maydeu-Olivares, A., & Joe, H. (2014). Assessing approximate fit in categorical data analysis. *Multivariate Behavioral Research, 49*(4), 305–328. https://doi.org/10.1080/00273171.2014.911075

Masyn, K. E. (2013). Latent class analysis and finite mixture modeling. In T. Little (Ed.), *The Oxford handbook of quantitative methods* (Vol. 2, pp. 551–611). Oxford University Press.

McLachlan, G., & Peel, D. (2000). *Finite Mixture Models*. John Wiley & Sons.

Meeusen, C., Meuleman, B., Abts, K., & Bergh, R. (2018). Comparing a variable-centered and a person-centered approach to the structure of prejudice. *Social Psychological and Personality Science, 9*(7), 767–777. https://doi.org/10.1177/1948550617720273

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13–103). Macmillan Publishing Co, Inc; American Council on Education.

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.

Morin, A. J. S., Boudrias, J.-S., Marsh, H. W., McInerney, D. M., Dagenais-Desmarais, V., Madore, I., & Litalien, D. (2017). Complementary variable- and person-centered approaches to the multidimensionality of psychometric constructs: Application to psychological well-being at work. *Journal of Business and Psychology, 32*(4), 395–419. https://doi.org/10.1007/s10869-016-9448-7

Muthén, L.K. and Muthén, B.O. (2013) Mplus (Version 7.11) [Computer Software]. Author, Los Angeles.

Nagin, D.S., 1999. Analyzing developmental trajectories: a semiparametric, group-based approach. *Psychol. Methods 4* (2), 139–157. https://doi.org/10.1037/1082- 989x.4.2.139

Nguyen, T. H., Han, H. R., Kim, M. T., & Chan, K. S. (2014). An introduction to item response theory for patient-reported outcome measurement. *Patient*, 7(1), 23–35. https://doi.org/10.1007/s40271-013-0041-0

Orlando, M., & Thissen, D. (2000). Likelihood-based item-fit indices for dichotomous item response theory models. *Applied Psychological Measurement, 24*(1), 50–64. https://doi.org/10.1177/01466216000241003

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

Piumatti, G., Cerutti, B., & Perron, N. J. (2021). Assessing communication skills during OSCE: need for integrated psychometric approaches. *BMC medical education*, *21*(1), 106. https://doi.org/10.1186/s12909-021-02552-8

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

R Core Team. (2025). *R: A language and environment for statistical computing* [Computer software]. R Foundation for Statistical Computing. https://www.R-project.org/

Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics*, *6*(2). https://doi.org/10.1214/aos/1176344136

Sorgente, A., Caliciuri, R., Robba, M., De Salve, F., & Lanz, M. (2025, September). Evaluating item quality from a person-centered perspective: Leveraging latent class analysis for item selection. Paper presented at the AIP Sezione Sperimentale Conference, Torino, Italy.

Sorgente, A., Caliciuri, R., Robba, M., Lanz, M., & Zumbo, B. D. (2025). A systematic review of latent class analysis in psychology: Examining the gap between guidelines and research practice. *Behavior Research Methods*, *57*(11), 301 https://doi.org/10.3758/s13428-025-02812-1

Sorgente, A., Lanz, M., Serido, J., Tagliabue, S., & Shim, S. (2019). Latent transition analysis: Guidelines and an application to emerging adults' social development. *TPM: Testing, Psychometrics, Methodology in Applied Psychology, 26*(1), 39-72. https://doi.org/10.4473/TPM26.1.3

Tong, L. K., Li, Y. Y., Liu, Y. B., Zheng, M. R., Fu, G. L., & Au, M. L. (2025). Validation of the short index of job satisfaction in Chinese nurses: classical test theory and item response theory. *International Journal of Nursing Studies Advances*, *8*. https://doi.org/10.1016/j.ijnsa.2025.100321

van Buuren, N., & Eggen, T. H. (2017). Latent-class-based item selection for computerized adaptive progress tests. *Journal of Computerized Adaptive Testing, 5*(2), 22-43 https://doi.org/10.7333/jcat.v5i2.62

Vuong Q. H. (1989). Likelihood ratio tests for model selection and non-nested hypotheses. *Econometrica: Journal of the Econometric Society*, 307–333. https://doi.org/10.2307/1912557

General Discussion and Conclusions

This dissertation was guided by two interconnected objectives: (a) to clarify the conceptual foundations of scientific reasoning, and (b) to provide cumulative validity evidence for its measurement. These aims converge toward a broader goal, building a coherent validity argument that justifies how test scores can meaningfully represent the construct they intend to measure. The project was grounded in the unified view of validity (Messick, 1989; Zumbo & Chan, 2014; AERA, APA, & NCME, 2014), which conceives validation not as a static outcome but as a progressive scientific argument. Within this framework, each piece of evidence contributes to the interpretive justification of scores rather than serving as an independent "type" of validity. To operationalize this approach, the dissertation followed the tripartite model originally proposed by Jane Loevinger (1957): the substantive, structural, and external phases of validation. This framework provided both the conceptual scaffolding and the methodological sequence that guided the three empirical studies: from defining the construct and its indicators (substantive), to evaluating its internal structure (structural), and finally to situating it within a network of theoretical and empirical relations (external).

The first study, an umbrella review, examined how scientific reasoning and its adjacent constructs – such as scientific inquiry, scientific literacy, and science process skills – have been conceptualized, measured, and validated across existing systematic reviews and meta-analyses. The synthesis revealed substantial conceptual ambiguity, wide heterogeneity in measurement instruments, and a general scarcity of psychometric reporting and external validation evidence. In doing so, the umbrella review provided the first higher-order synthesis of systematic reviews and meta-analyses on scientific reasoning, exposing not only the field's conceptual fragmentation but also its need for a unified framework capable of linking construct theory with psychometric practice. Building directly on these insights, the second study addressed these gaps by adapting and validating the Scientific Reasoning Scale (SRS) in the Italian context, thereby moving from conceptual clarification to empirical validation. Using a multi-method design that integrated back-translation, expert and cognitive interviews, and large-scale data collection (N = 897), this study provided new substantive, structural, and external evidence for the interpretation of SRS scores. Confirmatory Factor Analysis supported the unidimensional structure of the scale, while composite reliability indices confirmed its internal consistency; measurement invariance was demonstrated across gender, age, education, and ideological variables; and correlations with both cognitive (cognitive reflection, probabilistic reasoning) and belief-related constructs (climate change awareness and beliefs, paranormal health beliefs) supported its convergent and criterion-related validity. The study thus advanced the validation process from conceptual definition to empirical generalization, strengthening both theoretical coherence and psychometric rigor. To further investigate the sources of item-level weakness observed in the previous study, particularly the poor performance of the item assessing response bias, the third study deepened the analysis through a systematic comparison of measurement models. Specifically, it extended this work through a systematic comparison of measurement models – Classical Test Theory (CTT), Item Response Theory (IRT) – together with an exploratory Latent Class Analysis (LCA) used to evaluate item quality through additional fit and diagnostic indices. Each

framework offered a distinct analytical perspective: CTT provided indices of internal consistency and dimensionality; IRT estimated item difficulty and discrimination parameters; and LCA contributed exploratory evidence on item performance, complementing the structural evidence gathered through the other models. The complementary use of CTT and IRT enabled a fine-grained view of reliability and item functioning, while the introduction of LCA added a novel, person-centered diagnostic of item quality. Through this triangulation, the dissertation not only advanced the empirical understanding of the SRS but also modeled best practices in contemporary psychometrics, demonstrating how different analytical lenses can converge toward a single construct meaning.

These three studies form a cumulative validation trajectory that mirrors the very logic of the construct under investigation: iterative, evidence-based, and self-corrective. From conceptual clarification to empirical testing and methodological reflection, the dissertation embodies the principle that validity is not a property of a test, but a process of constructing scientific meaning. Through this integrative path, it advances both the theory of scientific reasoning and the practice of psychological measurement.

1. Advancing the theory of scientific reasoning

Research on scientific reasoning has long oscillated between domain-general and domain-specific accounts. Influential contributions have emphasized developmental inquiry skills and experimental problem solving (Kuhn, 1989; Klahr & Dunbar, 1988), epistemic cognition and evaluative stances toward knowledge (Kuhn, 2001), as well as interdisciplinary perspectives integrating reasoning and argumentation across educational contexts (Fischer et al., 2014; Fischer et al., 2018). At the same time, major conceptual reviews have documented substantial fragmentation in definitions and operationalizations of scientific reasoning, with partially overlapping constructs spanning cognitive, epistemic, and applied domains (Opitz et al., 2017; Díaz et al., 2021). Within this state of the art, the present dissertation advances the theory of scientific reasoning by proposing a deliberately circumscribed and method-centered conceptualization. Scientific reasoning is treated here as a domain-independent epistemic competence, applicable across natural sciences, social sciences, and humanities, insofar as it concerns shared principles for evaluating empirical claims. At the same time, this generality is not assumed to entail uniformity across all disciplinary practices. In line with Kind and Osborne's (2016) proposal of multiple "styles of scientific reasoning," the present framework acknowledges that different domains foreground distinct epistemic aims, entities, and standards of justification. Importantly, acknowledging the existence of multiple styles of scientific reasoning does not contradict the presence of a domain-general component. Rather, it clarifies that such generality resides in shared epistemic norms for evaluating evidence, which can be instantiated across disciplines despite differences in reasoning practices. On this basis, the present dissertation adopts a method-centered definition of scientific reasoning that focuses on principles of experimental and inferential evaluation—such as confounding control, bias detection, causal inference, and sensitivity to validity threats—that plausibly function as cross-domain epistemic norms.

This clarification allows the validity scope of the construct to be specified with precision. The conclusions drawn in this dissertation support the validity of Scientific Reasoning Scale (SRS) score inferences primarily for evidence-based evaluation of empirical claims, with strongest alignment to experimental and quasi-experimental logic. They are not intended to index all legitimate forms of scholarly reasoning, such as historical reconstruction or interpretive reasoning, which rely on partially distinct epistemic criteria. Explicitly articulating these scope conditions strengthens the theoretical coherence of the construct and enhances transparency in score interpretation, while also identifying a clear direction for future research: extending validation efforts to domain-specific contexts and examining whether interpretive stability is retained when scientific reasoning is operationalized through alternative disciplinary styles.

From a broader theoretical perspective, the findings across studies contribute to a richer understanding of scientific reasoning as a multidimensional and context-embedded competence. The construct emerges as simultaneously cognitive, epistemic, and civic. It involves not only the procedural skills of hypothesis testing, control of variables, and evidence evaluation, but also the epistemic dispositions that motivate individuals to seek justification, manage uncertainty, and coordinate theory with evidence in real-world contexts.

By integrating perspectives from psychology, education, and science communication, the dissertation extends the notion of reasoning scientifically beyond the laboratory. It conceptualizes scientific reasoning as an epistemic competence, a general capacity to regulate belief through evidence, and as a civic disposition enabling informed participation in societal debates. This view aligns with contemporary educational goals emphasizing scientific citizenship and lifelong learning.

At the empirical level, the analyses conducted on the Italian adaptation of the Scientific Reasoning Scale (SRS) offered valuable insight into how the construct materializes across its eleven facets. Each facet represents a distinct yet interrelated principle of sound scientific reasoning, ranging from recognizing blind and double-blind procedures, controlling for confounding variables, and ensuring construct and ecological validity, to understanding causality, maturation effects, reliability, and response bias. Together, these dimensions form a compact but comprehensive map of the reasoning processes underlying the evaluation of scientific evidence. Converging evidence from Classical Test Theory, Item Response Theory, and Latent Class Analysis revealed both strengths and vulnerabilities within this map. Some facets, such as those capturing the understanding of maturation effects, blind and double-blind procedures, confounding variables, and construct validity, consistently emerged as robust indicators of the underlying construct. They showed high loadings, strong discrimination, and clear contribution to latent-class differentiation, suggesting that they capture the epistemic core of causal reasoning. Other facets, including causality, control group, history, reliability, and ecological validity, displayed moderate but stable performance across analytical frameworks. Their psychometric adequacy, while acceptable, indicates that subtle refinements in item formulation could further enhance their diagnostic precision without altering theoretical coverage. More nuanced results emerged for the facets addressing random assignment and response bias. The former appeared particularly easy and showed partial non-invariance by gender, with women slightly more prone to incorrect responses at equivalent latent trait

levels. Although psychometrically sensitive, this facet remains theoretically indispensable for representing the logic of experimental control; hence, it should be retained and systematically monitored in future measurement invariance analyses. The latter facet, concerned with the recognition of response biases in data interpretation, proved the weakest across all approaches, with low loadings and minimal discrimination. Yet, because it represents a cornerstone of critical evaluation in scientific reasoning, its exclusion would impoverish construct representation. A targeted reformulation, rather than removal, is therefore recommended to preserve its conceptual centrality while improving psychometric performance.

These findings reaffirm the strength of the SRS as a theoretically grounded and psychometrically sound instrument. Despite localized weaknesses, the scale remains one of the few tools capable of assessing scientific reasoning across the general population, offering a clear operationalization of a complex competence. Continued refinement, particularly of the facets addressing response bias and experimental control, will enhance both its measurement precision and its theoretical reach.

Crucially, the validity of a measurement instrument depends not only on its internal coherence but also on the appropriateness of the inferences and uses it supports. In this sense, the Scientific Reasoning Scale should be interpreted with clear awareness of what its scores can, and cannot, tell us about individuals' reasoning competence. The present findings suggest that the Italian SRS is particularly informative in the lower to middle segments of the ability continuum, where it captures meaningful differences in how individuals understand and apply key scientific principles. Within this range, the scale allows for valid and useful inferences about the strengths and weaknesses of reasoning patterns, supporting its application in educational contexts, diagnostic assessments, and interventions aimed at fostering epistemic awareness and reflective thinking. At the same time, the analyses suggest a natural boundary to the interpretive reach of the scale. The SRS shows its greatest precision among individuals displaying emerging or intermediate levels of scientific reasoning, where it successfully captures meaningful variations in understanding and application. Its sensitivity gradually decreases at higher ability levels, where the items tend to provide less information. Rather than a limitation in itself, this pattern simply reflects the scale's current calibration. From a validity perspective, acknowledging this measurement focus is essential for maintaining appropriate score interpretations and for guiding future extensions aimed at broadening the scale's informational range. Future refinement efforts should therefore focus on extending the upper end of the scale's informational range, by developing more challenging items and expanding the item set within the existing facets, so as to balance measurement precision across the full continuum of ability. This would not only improve psychometric coverage but also enhance the theoretical representativeness of the construct, ensuring that the SRS can adequately reflect both the foundational and advanced forms of scientific reasoning as they manifest across diverse populations.

Complementary evidence from measurement invariance analyses further supports the interpretive robustness of the SRS. The scale operates equivalently across gender, age, and educational background, indicating that its items function similarly across diverse subgroups of the population. This ensures that observed differences can be meaningfully attributed to actual variations in scientific reasoning rather than to measurement artefacts.

Moreover, correlational and outcome-based evidence confirmed the expected pattern of relations with theoretically related constructs, such as epistemic beliefs, open-minded thinking, and trust in science, providing convergent and predictive support for the validity of score interpretations. Together, these findings justify cautious but confident use of the scale as an indicator of how individuals reason about evidence, causality, and validity.

Taken as a whole, the body of evidence accumulated across variable-centered (CTT, IRT) and person-centered (LCA) analyses points to a coherent and interpretable construct representation. The Italian SRS thus emerges as a concise yet theoretically rich instrument that captures essential facets of scientific reasoning in the general population. Its continued refinement, particularly through the inclusion of more challenging items and further validation across contexts, will strengthen both its psychometric precision and its contribution to understanding how people think scientifically in everyday life.

2. Advancing the theory of validity

Contemporary validity theory conceptualizes validity not as a property of tests, but as a property of the inferences drawn from test scores, supported by a cumulative body of evidence (Cronbach & Meehl, 1955; AERA, APA, & NCME, 2014; Messick, 1989). Despite broad endorsement of this unified framework, empirical practice in psychology often remains fragmented, with validation studies focusing on isolated psychometric indices rather than on coherent validity arguments (Flake et al., 2017; Hussey & Hughes, 2020).

Against this backdrop, the present dissertation advances validity theory in practice by explicitly structuring its empirical work around Loevinger's (1957) substantive, structural, and external components, and by treating validation as a cumulative, inference-centered process. This approach aligns with recent calls for greater transparency and integration in measurement practices (EFPA, 2025; Zumbo & Chan, 2014), and illustrates how contemporary validity theory can be operationalized in applied psychometric research.

The dissertation exemplifies how the contemporary view of validity can be translated into practice. Rather than treating content, structural, and external evidence as discrete "types," the research demonstrates how they can be accumulated as successive stages in a single interpretive argument. This operationalization of Loevinger's tripartite model thus becomes an empirical demonstration of cumulative validation in action.

The integration of qualitative and quantitative methods was central to this process. Expert evaluations and cognitive interviews provided response-process evidence, ensuring that items were semantically clear and culturally interpretable. The quantitative phases – CTT, IRT, and LCA – extended these insights by testing the internal logic of the scale from complementary perspectives. Together, they show how multiple sources of validity evidence can cohere into a single narrative of interpretive justification.

This triangulation of models also represents an innovative form of validity evidence not previously formalized: a methodological dialogue among frameworks that converge on construct meaning through different epistemic lenses. Such integration exemplifies the advantage of the contemporary validity paradigm, it increases

confidence in score inferences by cross-verifying the measurement argument from independent perspectives. In practical terms, it moves us closer to answering questions such as: “does a test score genuinely represent the intended construct, or might it reflect unrelated influences?”. Validity, in this light, is an epistemic safeguard, a means of ensuring that what we measure corresponds to what we claim to know.

This effort is far from merely technical. The consequences of using instruments without valid interpretive foundations reach deep into the ethical and social dimensions of psychology. When our inferences are not valid, our tests risk misrepresenting the individuals we aim to understand. In applied contexts—whether educational, clinical, or organizational—invalid inferences can reinforce stereotypes, sustain diagnostic stigma, or perpetuate structural inequalities under the guise of objectivity. As psychologists, our authority rests on the assumption that our measures *mean* what we say they mean. When this assumption is untested or unjustified, we do not only compromise the accuracy of our conclusions—we erode public trust in psychological science itself. Validity, therefore, is not just about methodological rigor: it is about scientific integrity and civic responsibility.

From this broader perspective, the contemporary movement toward fairness and inclusivity in testing represents a natural extension of the validity argument. Fairness is not an optional add-on to psychometric quality but a constitutive dimension of validity itself, because a measure that systematically disadvantages certain groups cannot be said to measure the intended construct faithfully. In recent years, scholars such as Bruno Zumbo have advanced what has been called a “fourth-generation” understanding of test fairness, one that recognizes how issues of validity and bias are inseparable from the cultural and social conditions in which testing takes place. This view moves beyond a merely statistical conception of bias and invites us to consider the intersections of identities, languages, and contexts that shape how people make sense of test content. In this light, concepts such as gender, education, or even scientific reasoning itself cannot be treated as fixed and universal categories. Their meanings are historically and culturally situated, constantly redefined by the social environments in which individuals live. A test that appears invariant across gender in one context may not be so in another, precisely because what it means to be “male” or “female,” or to engage in “scientific reasoning,” changes with culture, language, and time. Validity, therefore, is not achieved once and for all: it requires ongoing attentiveness to these intersections and to the voices of the diverse communities we seek to represent. Seen this way, validation becomes not only a methodological duty but an ethical commitment, to measure in ways that do not reproduce inequalities, to listen to how constructs resonate across cultures, and to ensure that the knowledge we produce about people is both accurate and just.

Finally, the dissertation aligns validity with the ethics of transparency. Through preregistration, open materials, and data sharing, it embodies the principles of open science as integral to the validation process itself. Methodological openness here is not merely procedural but epistemological: it transforms validation into a collective enterprise of trust, replication, and cumulative growth. Transparency in this sense serves a dual purpose. On the one hand, it strengthens the scientific credibility of our work by allowing others to scrutinize, reproduce, and extend our analyses. On the other, it fulfills a deeper civic function: it affirms psychology’s

commitment to accountability and public trust. Making our data, materials, and decisions visible is not simply an act of compliance with emerging norms, it is a declaration of intellectual honesty and humility. It acknowledges that psychological knowledge is provisional, contextual, and enriched by dialogue. When researchers share not only their successes but also their uncertainties and methodological constraints, they transform validation into a transparent conversation rather than a hidden procedure. This openness allows validity to grow cumulatively across studies and generations, preventing the isolation of single projects and fostering the integration of evidence within the broader scientific community. In this way, open science becomes a practical ethics of validity: a way of aligning the pursuit of accuracy with the values of fairness, reproducibility, and collective progress.

Ultimately, transparency closes the circle that begins with validity. To validate a measure is to claim that it speaks truthfully about people's thoughts, abilities, or experiences. To do so transparently is to recognize that such claims carry moral weight, and that credibility in science is inseparable from openness to examination. By embedding transparency at every stage, from preregistration to data sharing, this dissertation affirms that validity and openness are not separate ideals but two expressions of the same ethical commitment: to make our measures, our reasoning, and our science itself worthy of trust. In line with this view, Appendix B includes an open science workflow of the dissertation, which documents how transparency practices were implemented across all studies, demonstrating that openness can be systematically integrated into each phase of the research process.

3. Limitations as spaces for growth

No validation process is ever complete; each step clarifies the construct even as it exposes new questions. The present dissertation represents a foundation rather than an endpoint—a framework from which further theoretical, methodological, and practical developments can grow. The following sections outline four main areas for future work: (a) refinements in reliability and content coverage, (b) theoretical and methodological extensions of the construct, (c) ecological and developmental generalization, and (d) ethical and collaborative dimensions of validation.

Reliability and Content Coverage. Although the scale demonstrated adequate internal consistency, reliability coefficients were relatively modest, reflecting both the heterogeneity of item content and the brevity of the measure. One potential avenue for improvement would be to develop an extended version of the scale in which each facet of scientific reasoning is represented by multiple items. Such an expansion could enhance internal consistency while also broadening the range of item difficulty within each facet, thereby improving measurement precision across different ability levels. A longer form would not alter the construct but would capture it with greater granularity, offering richer data for both psychometric and substantive analyses.

In addition, one item addressing a core facet of scientific reasoning, response bias, was removed due to weak psychometric performance, which slightly reduced content coverage. This removal, however, is not merely a technical adjustment but a theoretical signal. Each facet of the SRS represents a specific way in which

individuals coordinate evidence and inference; omitting one means temporarily narrowing the conceptual map of scientific reasoning. The weak performance of the response-bias item suggests not that the facet is irrelevant, but that the item failed to represent it adequately within the Italian context, perhaps due to linguistic nuances or cultural interpretations of the scenario.

Future work should therefore aim to reformulate and reintroduce this facet, ensuring that it captures the ability to detect distortions in data interpretation, an essential aspect of epistemic vigilance and critical evaluation. Revisiting this component will not only restore full coverage of the construct but also advance its theoretical coherence, clarifying how sensitivity to response bias interacts with other principles such as control of variables, causal inference, and evidence evaluation. In this sense, psychometric refinement becomes a vehicle for theoretical progress: each revision contributes to a more complete and precise understanding of what it means to reason scientifically.

Theoretical and Methodological Extensions. Beyond these item-level refinements, broader theoretical challenges remain. One persistent question concerns the interplay among the cognitive, epistemic, and social dimensions of scientific reasoning. While growing consensus supports its multidimensional nature, theoretical models still differ in how these components are weighted and integrated. Future research should therefore continue to articulate whether they represent distinct facets or unified expressions of a broader epistemic competence.

One promising direction for addressing this issue lies in the application of network psychometrics. This framework conceptualizes psychological constructs not as latent variables causing item responses, but as systems of mutually interacting components. Within this approach, each facet or item can be represented as a node in a network, and the strength of their interconnections reveals how elements of scientific reasoning co-activate and sustain one another. Applying network analysis to the SRS could thus advance the field on a double front. On the one hand, it would enrich theoretical understanding by mapping how scientific reasoning relates to adjacent constructs and by situating these within a broader, differentiated framework of epistemic competence. On the other hand, it would extend the structural validation of the SRS itself, offering a complementary lens to those provided by Classical Test Theory, Item Response Theory, and Latent Class Analysis.

While network psychometrics can deepen our structural understanding, cross-cultural validation can expand its contextual generalizability. Establishing measurement invariance across languages, cultures, and educational systems will help determine whether the principles of scientific reasoning are universal competences or culturally embedded epistemic practices. Such studies will not only strengthen the interpretive stability of the scale but also reveal culturally specific expressions of reasoning that can refine and expand the construct itself.

Ecological and Developmental Generalization. Another important consideration concerns the ecological generalizability of the current measure. The SRS relies on brief, laboratory-style scenarios that capture the

formal logic of experimental reasoning but only partially represent how people engage with evidence in their everyday environments. Real-world reasoning is often situated, uncertain, and socially mediated, it involves evaluating claims in the media, interpreting health information, or judging the credibility of sources in daily decisions. These contexts demand the same epistemic skills assessed by the SRS, yet they unfold in less structured and more ambiguous conditions.

To address this gap, future research should explore daily or context-embedded versions of scientific reasoning assessments, capable of capturing how individuals apply scientific principles outside of formal experimental settings. In this direction, the present research program has already initiated the development of a daily version of the SRS, designed to assess reasoning processes as they occur in naturalistic contexts. Preliminary analyses on this pilot version indicated partial factorial coherence, suggesting that additional item refinement will be required to achieve stable measurement across diverse situations. While exploratory, this effort points toward a broader goal: enhancing the ecological validity and generalizability of scientific reasoning measures. The materials, items, and preliminary analyses of the daily version are openly available on OSF, offering a transparent basis for future collaborative refinement.

Complementing these ecological perspectives, longitudinal research can illuminate how scientific reasoning evolves across time and educational transitions. Tracking reasoning trajectories would clarify how individuals acquire, consolidate, or even lose epistemic skills over the lifespan. These studies would not only strengthen the evidence base for test–retest validity but also inform theoretical models of epistemic growth, linking scientific reasoning to learning processes, metacognitive regulation, and exposure to scientific discourse.

Consequential and Ethical Dimensions. Finally, future work should also address consequential validity—the often-overlooked dimension concerning the real-world implications of how test scores are interpreted and used. Beyond assessing direct test effects, this perspective examines whether the interpretations derived from the SRS lead to educationally and civically sound outcomes. Do assessments of scientific reasoning foster metacognitive awareness and reflection about one's own thinking? Do they cultivate skepticism toward misinformation, strengthen respect for evidence, or encourage constructive participation in public debates on science-related issues? Considering these broader consequences situates the SRS within the civic mission of psychological measurement: to support the development of reflective, informed, and responsible citizens.

Throughout these developments, the principles of open science remain indispensable. Preregistration, shared data and code, and collaborative replication efforts are not merely procedural ideals but the infrastructure through which cumulative validity can grow. There is still much work to be done, the present dissertation represents only a starting point, but with all materials openly available, anyone can join, replicate, or extend this work, contributing to a collective effort to advance the measurement and understanding of scientific reasoning.

4. Epilogue – The ethics of reasoning

Beyond results and analyses, what remains is the ethical gesture that guided this work. Every scientific journey begins with curiosity, but it matures through responsibility. This dissertation began with a simple intuition: science is not confined to laboratories or formal instruction, but lives quietly in the gestures of everyday life. I return here to that initial image: my grandfather in his vineyard, tending the vines with patient precision, observing, adjusting, repeating. His was not academic science, but a form of reasoning guided by observation, hypothesis, and care. It embodied the same logic that underlies this work: to look closely, to test one's understanding, and to learn from the dialogue between expectation and evidence.

In many ways, the process of validation mirrors that gesture. Like the vintner who must trust both method and intuition, validation demands attentiveness, humility, and perseverance. It is not a mechanical verification, but a discipline of responsibility—an ethical commitment to transparency, accuracy, and coherence. Each analysis and revision in this research has been an act of observation and adjustment, part of a larger cycle aimed at refining both the instrument and the understanding it seeks to capture.

At its core, this work carries three messages. First, scientific reasoning is a shared and democratic competence, an inheritance of our humanity rather than a privilege of expertise. It reminds us that the capacity to weigh evidence, question assumptions, and tolerate uncertainty is foundational to both personal judgment and collective life. Second, validation is a process of cumulative knowledge, a way of constructing meaning that honors the provisional and self-correcting nature of science. By integrating theory, method, and evidence across studies, this dissertation has sought to model how knowledge in psychology can grow through coherence rather than closure. Third, psychological research is, at its best, a dialogue between method and humanity: it aspires to measure not to reduce, but to illuminate. Measurement becomes meaningful only when it remains aware of the human reasoning it seeks to understand.

Ultimately, the spirit of this work lies in a simple conviction: that science is a form of care—for truth, for precision, and for the human capacity to make sense of the world. The scientific reasoning that guides researchers in their analyses is the same reasoning that guides ordinary people as they interpret experience, decide, doubt, and learn. To study it, therefore, is to study something deeply human.

And so this dissertation closes as it began, with a return to light and method. The task of research, like the tending of a vineyard, is never finished; it is a cycle of attention and renewal. Our responsibility as scholars is to illuminate with method what already lives in us as curiosity and reason—to make visible, through science, the quiet intelligence of being human.

Bibliography

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

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin, 52*(4), 281–302. https://doi.org/10.1037/h0040957

Díaz, C., Dorner, B., Hussmann, H., & Strijbos, J.-W. (2021). Conceptual review on scientific reasoning and scientific thinking. *Current Psychology: A Journal for Diverse Perspectives on Diverse Psychological Issues.* Advance online publication. https://doi.org/10.1007/s12144-021-01786-5

European Federation of Psychologists' Associations AISBL (EFPA, 2025). Model for the Review, Description and Evaluation of Psychological and Educational Tests (Test Review Model): https://www.efpa.eu/resources

Fischer, F., Kollar, I., Ufer, S., Sodian, B., Hussmann, H., Pekrun, R., Neuhaus, B., Dorner, B., Pankofer, S., Fischer, M., Strijbos, J.-W., Heene, M., & 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

Fischer, F., Chinn, C. A., Engelmann, K., & Osborne, J. (Eds.). (2018). *Scientific reasoning and argumentation: The roles of domain-specific and domain-general knowledge*. Routledge. https://doi.org/10.4324/9780203731826

Flake, J. K., Pek, J., & Hehman, E. (2017). Construct validation in social and personality research: Current practice and recommendations. *Social Psychological and Personality Science, 8*(4), 370–378. https://doi.org/10.1177/1948550617693063

Hussey, I., & Hughes, S. (2020). *Hidden invalidity among 15 commonly used measures in social and personality psychology. Advances in Methods and Practices in Psychological Science, 3*(2), 166–184. https://doi.org/10.1177/2515245919882903

Kind, P. E. R., & Osborne, J. (2017). Styles of scientific reasoning: A cultural rationale for science education?. Science education, 101(1), 8-31. https://doi.org/10.1002/sce.21251

Klahr, D., & Dunbar, K. (1988). Dual space search during scientific reasoning. *Cognitive science*, *12*(1), 1-48.

Kuhn, D. (1989). Children and adults as intuitive scientists. *Psychological review*, *96*(4), 674.

Kuhn, D. (2001). How Do People Know? *Psychological Science, 12*(1), 1–7. https://doi.org/10.1111/1467-9280.00302

Loevinger, J. (1957). Objective tests as instruments of psychological theory. *Psychological Reports, 3*(3), 635–694. https://doi.org/10.2466/pr0.1957.3.3.635

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13–103). American Council on Education & Macmillan.

Zumbo, B. D., & Chan, E. K. H. (Eds.). (2014). *Validity and validation in social, behavioral, and health sciences.* Springer International Publishing/Springer Nature. https://doi.org/10.1007/978-3-319-07794-9

Appendix B – Open Science Workflow of the Dissertation

This dissertation was conceived and conducted in full adherence to the principles of Open Science and research transparency. Each empirical study was preregistered, and all materials, datasets, and analysis scripts are openly accessible on the Open Science Framework (OSF). These practices were implemented to ensure reproducibility, accountability, and cumulative progress in the study of scientific reasoning.

Open Science Practices Across the Dissertation

Stage of the Research Process	Implementation and Open Resources
1. Concept & Design	Study planning and construct definition were guided by transparent reporting standards (AERA, APA, & NCME, 2014) and by the EFPA Test Review Model (EFPA, 2025), which emphasize explicit documentation, fairness evaluation, and openness throughout test adaptation and validation.
2. Preregistration	All empirical studies were preregistered on OSF prior to data collection, including hypotheses, variables, and analytic plans: *Umbrella Review (chapter 3)*: https://doi.org/10.17605/OSF.IO/U4STH *Italian Adaptation/Validation and Measurement-Model Comparison for the SRS (chapters 4 and 5):* - https://doi.org/10.17605/OSF.IO/AYZX2; - https://doi.org/10.17605/OSF.IO/HR75K
3. Data Collection	Procedures, instruments, and consent forms were openly documented. Ethical approvals were obtained and uploaded to OSF where applicable.
4. Analysis & Reproducibility	Analysis scripts (SPSS, R, Mplus) were shared to ensure reproducibility. All syntax and output files are stored in their respective OSF repositories, covering all the analyses conducted.
5. Open Materials & Data	Anonymized datasets and instrument files are shared under a CC BY 4.0 license. Materials and scoring keys are available for reuse in further studies.
6. Transparent Publication	Each empirical manuscript was linked to its OSF repository, ensuring open access to supplementary materials and reproducible workflows.

7. Cumulative Science	The open documentation of data, materials, and validation processes was designed to facilitate meta-analytic integration and the cumulative refinement of measurement validity arguments.

Notes on Preregistrations and Research Developments

While all empirical studies were preregistered on the Open Science Framework, the research process evolved in ways that extended beyond the initial plans, reflecting the iterative and self-corrective nature of validation.

The umbrella review followed its preregistered protocol closely, including the search strategy, inclusion criteria, and coding framework. For the Italian adaptation and validation studies, the preregistration initially planned to gather validity evidence for the Scientific Reasoning Scale through both Classical Test Theory and Item Response Theory. However, as detailed in Chapter 4, preliminary analyses revealed conceptual and psychometric issues that required revising and adapting the scale before further validation. Consequently, a second data collection was carried out with the revised version, which became the foundation for the analyses reported in the same chapter. The comparison of measurement models presented in Chapter 5 was based on this second dataset.

Notably, the inclusion of Latent Class Analysis was *not part of the preregistered plan* but emerged as a later methodological extension to explore item functioning from a person-centered perspective.

Finally, one of the preregistrations also described the development of a daily version of the SRS, designed to capture short-term fluctuations in reasoning across everyday contexts. Although this version was developed and administered, preliminary analyses indicated psychometric limitations. For this reason, the daily SRS was not included in the present dissertation and will be further refined in future research.

This transparent documentation of deviations and unplanned developments aligns with the principles of Open Science and with the logic of cumulative validation that guided this project—acknowledging that refinement, revision, and reanalysis are integral parts of the scientific reasoning process itself.

Open Science was not only a set of technical practices but a guiding epistemic stance throughout this dissertation. Transparency, reproducibility, and community accessibility were treated as essential components of the validity argument and as expressions of scientific reasoning itself.

La borsa di dottorato è stata cofinanziata con risorse del Piano Nazionale di Ripresa e Resilienza PNRR Componente 1 “Potenziamento dell’offerta dei servizi di istruzione: dagli asili nido all’Università” - Componente 2 “Dalla Ricerca all'Impresa”, Investimento 3.3, Introduzione di dottorati innovativi che rispondono ai fabbisogni di innovazione delle imprese e promuovono l'assunzione dei ricercatori da parte delle imprese”