The Eurasia Proceedings of
Educational & Social Sciences (EPESS)
ISSN: 2587-1730

The Eurasia Proceedings of Educational and Social Sciences (EPESS), 2025

Volume 45, Pages 88-97

ICRET 2025: International Conference on Research in Education and Technology

# Assessing Scientific Thinking in Early Childhood: Development and Validation of the Scientific Thinking Skills Assessment Tool (STS-AT)

**Elif Ozturk**
Indiana University of Bloomington
Giresun University

**Abstract** This study reports the development and validation of the Scientific Thinking Skills Assessment Tool (STS-AT), designed to measure scientific thinking in children aged 5–8 years. The STS-AT assesses four domains: critical inquiry, hypothesis testing, analytical interpretation, and metacognitive awareness. Item development was guided by theoretical frameworks and expert review, followed by pilot testing with 72 children, which demonstrated clarity and inter-rater reliability (Cohen's $\kappa = 0.88$). The final instrument comprised 12 open-ended tasks supported with visual aids and scored on a four-point rubric. The main study involved 282 children from Turkish kindergartens and primary schools. Reliability analyses indicated strong internal consistency (Cronbach's $\alpha = .87$) and high test–retest stability ($r = .91$). Exploratory and confirmatory factor analyses supported a four-factor structure with excellent fit (RMSEA = .04, CFI = .95, TLI = .93, SRMR = .06). Results showed significant improvements in scientific thinking with age ($F(3,278) = 18.81$, $p < .001$, $\eta^2 = 0.17$), while no gender differences were observed ($t = -1.01$, $p = 0.315$). These findings suggest that the STS-AT is a valid, reliable, and developmentally appropriate tool for assessing scientific thinking in early childhood.

**Keywords:** Scientific thinking, Early childhood, Assessment, Scale development, Validity, Reliability

## Introduction

Scientific thinking has long been recognized as a cornerstone of cognitive development in early childhood, encompassing the skills of questioning, predicting, hypothesizing, testing ideas, interpreting evidence, and reflecting on one's reasoning (Kuhn, 2010; Zimmerman, 2007). Classic developmental theories position this period as critical: Piaget (1972) described the transition from preoperational to concrete operational thought as a time when children increasingly coordinate evidence and reasoning, while Vygotsky (1978) emphasized the role of social interaction and scaffolding in fostering inquiry and reflection.

In recent decades, researchers have conceptualized young children as "little scientists" who actively generate and test explanations about the natural and social world (Gopnik et al., 2000). More recent empirical work confirms that children as young as five can design simple experiments, differentiate between confounded and unconfounded evidence, and modify their explanations based on feedback (Koerber et al., 2015; Köksal, 2022). These findings highlight not only the presence of early competencies but also the importance of providing structured opportunities to nurture them.

The global emphasis on STEM education has further reinforced the importance of fostering scientific thinking in early years (Bybee, 2013; OECD, 2017). Early scientific reasoning is associated with later academic achievement, problem-solving, and civic scientific literacy (National Research Council [NRC], 2012). Moreover, cultivating inquiry and reflective skills in childhood contributes to the development of critical 21st-century competencies such as creativity, resilience, and informed decision-making (NGSS Lead States, 2013).

Despite this recognition, assessment practices in early childhood education remain limited. Existing tools often measure isolated skills—such as observation or prediction—rather than capturing the multidimensional nature

of scientific thinking (Zimmerman, 2007; Koerber & Osterhaus, 2020). The Work Sampling System (WSS) includes a "scientific thinking" domain but does not provide fine-grained psychometric evidence and treats science as a subset of general learning (Meisels et al., 1995). More recently, Koerber and Osterhaus (2020) developed the Science-K Inventory**,** a Rasch-scaled assessment for preschoolers, focusing on experimentation and nature-of-science concepts. While valuable, such instruments remain relatively rare and often lack integration of metacognitive elements—such as children's awareness of their own thinking—which are known to be crucial even in early development (Flavell, 1979; Schraw & Dennison, 1994).

Therefore, there is a pressing need for a valid and reliable instrument that conceptualizes scientific thinking holistically, integrating critical inquiry**,** hypothesis testing**,** analytical reasoning**,** and metacognitive awareness in a developmentally appropriate framework. Addressing this gap, the present study introduces the Scientific Thinking Skills Assessment Tool (STS-AT), specifically designed for children aged 5–8 years. Drawing on constructivist theories and contemporary evidence, the STS-AT provides a child-centered, play-based, and psychometrically robust measure.

The purpose of this paper is to describe the conceptualization, development, and validation of the STS-AT. We present evidence of internal consistency, test–retest reliability, item-level performance, and factorial validity (via exploratory and confirmatory factor analyses). Additionally, developmental validity is examined through age comparisons, and potential gender differences are explored. By establishing a rigorous foundation, this study aims to contribute a novel and reliable tool to the field of early childhood science education.

## Theoretıcal Background

Scientific thinking in early childhood is a multidimensional construct that emerges through the interaction of cognitive development, social context, and instructional opportunities. The STS-AT was designed to reflect four interrelated domains—critical inquiry, hypothesis testing, analytical interpretation, and metacognitive awareness—each grounded in established theoretical and empirical literature.

### Critical Inquiry

Critical inquiry refers to children's capacity to generate questions, attend to observations, and identify meaningful problems for investigation. Piaget (1972) emphasized that during the transition from preoperational to concrete operational stages, children begin to coordinate their observations with logical operations. Vygotsky's (1978) sociocultural theory further highlighted that inquiry skills are fostered when children are guided by more knowledgeable peers or adults within the zone of proximal development. Contemporary research shows that preschool and early primary students are capable of posing causal questions and noticing patterns in phenomena when supported with scaffolding (Eshach & Fried, 2005; Köksal, 2022). Inquiry is also positioned as a central scientific practice in the Next Generation Science Standards (NGSS Lead States, 2013), underlining its educational relevance.

### Hypothesis Testing

The ability to generate predictions and verify them through observation or experimentation is central to scientific reasoning. Kuhn (2010) describes hypothesis testing as a critical shift from intuitive explanations to evidence-based thinking. Empirical studies demonstrate that children as young as five can engage in simple experimental designs and revise their hypotheses in light of outcomes (Koerber et al., 2015). According to Zimmerman (2007), children's competence in controlling variables and recognizing unconfounded evidence increases markedly between ages 5 and 8. The National Research Council (2012) also identifies prediction and verification as key practices in developing scientific literacy from early schooling.

### Analytical Interpretation

Analytical interpretation involves drawing inferences, recognizing cause–effect relations, and applying logical reasoning to data. Research shows that children develop the ability to distinguish correlation from causation during the early school years, although scaffolding is often needed (Sodian et al., 1991). Koerber and Osterhaus (2020) argue that analytic reasoning is a separate yet related dimension of scientific thinking, requiring both

domain-general skills and specific knowledge. Developmental psychology suggests that such reasoning is not merely about generating correct answers but about cultivating explanatory frameworks that integrate evidence and logic (Zimmerman, 2007).

### Metacognitive Awareness

Metacognitive awareness is the ability to reflect on and regulate one's own thinking. Flavell (1979) introduced metacognition as a crucial developmental process, while Schraw and Dennison (1994) demonstrated that even young learners exhibit early forms of metacognitive awareness when asked to evaluate their understanding. Recent studies confirm that children can monitor their confidence, recognize uncertainty, and adjust strategies accordingly (Kuhn, 2000; Whitebread et al., 2009). Incorporating metacognition into assessments provides richer insight into children's scientific reasoning, as it captures not only what they know but how they know it.

Together, these four domains reflect a comprehensive approach to scientific thinking in early childhood. By grounding the STS-AT in both classic developmental theories and contemporary frameworks, the instrument ensures ecological and educational validity. Moreover, this multidimensional model aligns with international policy calls for integrating inquiry, reasoning, and reflection into early STEM education (OECD, 2017; NGSS Lead States, 2013).

## Method

### Participants

The study was conducted in two phases: a pilot study and a main validation study. In the pilot study, seventy-two children (35 girls, 37 boys; *M* age = 6.4 years) from two public kindergartens in Türkiye participated. The pilot aimed to examine item clarity, engagement, and scoring feasibility. The main study included 282 children (138 girls, 144 boys) aged between 5 and 8 years, recruited from both urban and rural schools in northern Türkiye. Parental consent and children's assent were obtained prior to participation. The distribution of participants across age groups is presented in Table 1.

Table 1. Demographic characteristics of the main study sample

| Age (years) | Girls (n) | Boys (n) | Total (n) | Mean Age (SD) |
|---|---|---|---|---|
| 5 | 34 | 37 | 71 | 5.2 (0.4) |
| 6 | 48 | 50 | 98 | 6.3 (0.5) |
| 7 | 34 | 33 | 67 | 7.2 (0.4) |
| 8 | 22 | 24 | 46 | 8.1 (0.5) |
| Total | 138 | 144 | 282 | 6.5 (1.1) |

### Instrument Development

The Scientific Thinking Skills Assessment Tool (STS-AT) was developed to capture four theoretically and empirically grounded domains: critical inquiry, hypothesis testing, analytical interpretation, and metacognitive awareness. Item generation was informed by constructivist and sociocultural theories (Piaget, 1972; Vygotsky, 1978) as well as empirical literature on scientific reasoning in early childhood (Zimmerman, 2007; Koerber et al., 2015). Twelve open-ended items were created, with three items representing each domain. The items were reviewed by eight experts in early childhood education, developmental psychology, and science education using a Delphi procedure. The Content Validity Index (CVI) across items was .91, indicating high agreement regarding the relevance and clarity of the items.

Following expert review, the instrument was piloted with seventy-two children. Pilot results supported both feasibility and reliability, with inter-rater agreement between two independent coders reaching $\kappa = .88$. Based on pilot feedback, items were revised to enhance concreteness and engagement. For instance, abstract prompts such as "What do you think happens next?" were replaced with developmentally appropriate tasks like "What happens if we add more blocks to the tower?" Table 2 presents example items from the STS-AT along with the associated scoring rubric.

Table 2. Example Items from the STS-AT

| Domain | Example Item | Scoring Rubric (1–4) |
|---|---|---|
| Critical Inquiry | What questions would you ask if you found a new bug? | 1 = vague → 4 = specific scientific question |
| Hypothesis Testing | What do you think will happen if we put the paper boat in water? | 1 = no prediction → 4 = clear testable prediction |
| Analytical Interpretation | Why do you think the block tower fell down? | 1 = irrelevant → 4 = logical causal explanation |
| Metacognitive Awareness | How did you decide your answer? | 1 = no reflection → 4 = explicit self-reflection |

**Procedure**

The STS-AT was administered individually in quiet classroom settings by trained facilitators. Each session lasted approximately twenty minutes per child. Standardized administration protocols were followed, including scripted instructions, visual prompts, and scoring guidelines, to minimize assessor bias. To increase accessibility, visual aids and manipulatives such as blocks, paper boats, and picture cards were used to scaffold children's responses. Ethical approval was obtained from the university's ethics committee (Ref. No. 2025/04-12). Written parental consent and verbal assent from the children were required for participation. Inter-rater reliability was reassessed in a randomly selected 25% of the main sample, yielding strong agreement ($\kappa = .91$).

**Data Analysis**

Analyses were conducted using SPSS 29 and AMOS 27 and followed established guidelines for scale development (DeVellis, 2017; Tabachnick & Fidell, 2019). Item analysis included computation of means, standard deviations, corrected item–total correlations, and Cronbach's alpha if item deleted. Reliability evidence included Cronbach's alpha coefficients for the total scale and subscales, inter-rater reliability, and test–retest reliability over a two-week interval with a subsample of fifty children.

Construct validity was examined in two stages. First, exploratory factor analysis (EFA) with principal axis factoring and oblique rotation was conducted to explore underlying factor structure. Second, confirmatory factor analysis (CFA) using maximum likelihood estimation tested the hypothesized four-factor model, with model fit evaluated using $\chi^2/df$, RMSEA, CFI, TLI, and SRMR. Convergent and discriminant validity were examined through intercorrelations among the subscales.

Developmental validity was evaluated by comparing total STS-AT scores across age groups using one-way ANOVA. Post hoc analyses were performed to identify significant group differences, with eta-squared ($\eta^2$) reported as an effect size measure. Gender comparisons were conducted using independent samples *t*-tests, with Cohen's *d* reported to indicate the magnitude of differences. Table 3 summarizes the analytic procedures used in this study.

Table 3 Overview of analytic procedures

| Analysis Type | Purpose | Indicators Reported |
|---|---|---|
| Item Analysis | Evaluate item quality | Mean, SD, corrected r, α if deleted |
| Reliability | Internal consistency and stability | Cronbach's α, test–retest, κ |
| Construct Validity (EFA, CFA) | Confirm factor structure | Variance explained, loadings, fit indices |
| Developmental Validity | Age-related progression | ANOVA, $\eta^2$, post hoc contrasts |
| Gender Differences | Gender fairness | t-test, Cohen's d |

Following the analytic framework summarized in Table 3, the study applied a systematic and multi-step validation process to establish the psychometric quality of the STS-AT. Item analyses ensured that each task functioned as intended, while reliability testing provided evidence of both internal consistency and temporal stability. The combination of exploratory and confirmatory factor analyses allowed for a robust evaluation of construct validity, confirming the four-domain model hypothesized on theoretical grounds. Finally, developmental and gender-based comparisons offered additional insights into the sensitivity and fairness of the instrument across subgroups. Together, these analyses created a comprehensive foundation for interpreting the results reported in the following section.

## Results

### Descriptive Statistics

The initial analysis focused on descriptive statistics of the STS-AT. Across the main sample ($N = 282$), the mean total score was 24.82 ($SD = 4.91$), indicating a moderate level of scientific thinking skills in children aged 5–8 years. Examination of the four domains revealed comparable distributions, although some variation was observed in the relative strengths of subdomains.

Table 4 presents the descriptive statistics and Cronbach's alpha reliability coefficients for each subscale. Analytical interpretation showed the highest mean score, suggesting that children were relatively adept at recognizing cause–effect relationships and drawing logical inferences. In contrast, metacognitive awareness yielded the lowest mean, consistent with the notion that reflective thinking develops later than direct reasoning skills.

Table 4 Descriptive statistics and reliability of STS-AT Subscales

| Subscale | Mean | SD | Cronbach's α |
|---|---|---|---|
| Critical Inquiry | 6.48 | 1.80 | .78 |
| Hypothesis Testing | 5.99 | 1.90 | .79 |
| Analytical Interpretation | 6.65 | 1.70 | .82 |
| Metacognitive Awareness | 5.72 | 1.60 | .79 |
| Total Scale | 24.82 | 4.91 | .87 |

The total scale reliability coefficient of α = .87 suggests strong internal consistency. Overall, the descriptive statistics support the internal coherence of the instrument and provide preliminary evidence that the four domains function as theoretically expected.

### Item Analysis

Item-level analysis was conducted to evaluate the functioning of each of the twelve items. Mean scores ranged between 1.98 and 2.41, demonstrating variability across items and indicating that the tasks provided an appropriate level of challenge for children. Corrected item–total correlations varied between .35 and .52, all exceeding the recommended threshold of .30. This suggests that each item contributed meaningfully to the construct being measured. Cronbach's alpha if item deleted ranged from .85 to .87, showing that no single item significantly weakened the overall reliability of the scale. Table 5 summarizes these findings in detail.

Table 5. Item means, standard deviations, corrected ıtem–total correlations, and alpha if deleted

| Item | Mean | SD | Corrected r | α if deleted |
|---|---|---|---|---|
| ESM1 | 2.15 | 0.82 | .38 | .86 |
| ESM2 | 2.07 | 0.79 | .41 | .85 |
| ESM3 | 2.23 | 0.81 | .42 | .85 |
| H1 | 2.31 | 0.84 | .35 | .86 |
| H2 | 1.98 | 0.77 | .37 | .86 |
| H3 | 2.16 | 0.80 | .40 | .85 |
| AY1 | 2.39 | 0.85 | .52 | .85 |
| AY2 | 2.41 | 0.82 | .47 | .85 |
| AY3 | 2.28 | 0.80 | .45 | .85 |
| MF1 | 2.04 | 0.79 | .36 | .86 |
| MF2 | 2.11 | 0.77 | .39 | .85 |
| MF3 | 2.08 | 0.81 | .40 | .85 |

The overall pattern suggests that items were well balanced in terms of difficulty and discrimination. Importantly, the relatively higher correlations for analytical interpretation items (AY1–AY3) reinforce the robustness of this subscale.

### Reliability

Reliability analyses provided strong evidence of measurement consistency. The total scale demonstrated high internal consistency (α = .87), while subscale reliabilities ranged from .78 to .82. These values are well above the commonly accepted threshold of .70. Test–retest reliability assessed with fifty children over a two-week interval yielded $r$ = .91, confirming excellent temporal stability. Inter-rater reliability, assessed in 25% of randomly selected cases, also demonstrated high agreement (κ = .91), underscoring the robustness of scoring procedures. Taken together, these findings support the reliability of the STS-AT across different raters and occasions.

### Factor Analyses

Construct validity was examined through factor analytic techniques. Exploratory factor analysis (EFA) supported a four-factor solution consistent with the hypothesized domains, explaining 65% of the total variance. All items loaded strongly on their intended factors, with loadings above .59. Confirmatory factor analysis (CFA) further tested the four-factor model and yielded excellent fit indices, $\chi^2/df = 1.92$, RMSEA = .04, CFI = .95, TLI = .93, SRMR = .06. These results provide strong evidence that the STS-AT captures a multidimensional construct aligned with theoretical expectations.

### Age Differences

Developmental validity was evaluated by comparing children's scores across age groups. A one-way ANOVA revealed significant differences, $F(3, 278) = 18.81$, $p < .001$, $\eta^2 = .21$. Table 3 presents the group means and standard deviations.

Table 6. Total STS-AT scores by age group

| Age (years) | n | Mean | SD |
|---|---|---|---|
| 5 | 71 | 22.18 | 4.32 |
| 6 | 98 | 26.19 | 4.89 |
| 7 | 67 | 26.27 | 5.01 |
| 8 | 46 | 27.19 | 4.85 |

The results indicate a clear developmental trend, with older children outperforming younger ones. Notably, five-year-olds scored significantly lower than the older groups, while the differences between seven- and eight-year-olds were minimal. This pattern is consistent with developmental theories that predict rapid gains in reasoning and problem-solving between ages five and seven, followed by consolidation at later ages.

### Gender Differences

Gender comparisons showed that girls (M = 25.07, SD = 4.80) scored slightly higher than boys (M = 24.59, SD = 5.02), although this difference was not statistically significant, $t(278) = –1.01$, $p = .315$, $d = –0.12$. Table 4 presents these findings.

Table 7. Comparison of STS-AT scores by gender

| Gender | n | Mean | SD |
|---|---|---|---|
| Girls | 138 | 25.07 | 4.80 |
| Boys | 144 | 24.59 | 5.02 |

The absence of significant gender differences suggests that the STS-AT is free from bias and performs equivalently across boys and girls. This finding also aligns with contemporary research showing that gender gaps in early scientific reasoning are minimal when children are provided with similar opportunities.

## Discussion

The present study introduced and validated the Scientific Thinking Skills Assessment Tool (STS-AT) for children aged 5–8 years, providing multi-source evidence for reliability and construct validity. Internal consistency for the total scale (α = .87) and subscales (α = .78–.82), excellent inter-rater agreement, and high short-term stability collectively support the score reliability of the instrument, consistent with best practices

articulated in the *Standards for Educational and Psychological Testing* (AERA/APA/NCME, 2014). The factor-analytic results further substantiate a theoretically coherent, four-factor structure, with EFA indicating substantial explained variance and CFA reflecting excellent global fit ($\chi^2/df = 1.92$, RMSEA = .04, CFI = .95, TLI = .93, SRMR = .06). These indices are well within widely cited benchmarks for acceptable to good fit (e.g., RMSEA ≤ .06, CFI/TLI ≥ .95, SRMR ≤ .08), which strengthens the argument that STS-AT captures interrelated but distinct dimensions of early scientific thinking.

Interpreted against the developmental literature, the observed age-related gains on the STS-AT are theoretically expected and empirically consonant with prior work. Extensive reviews and large-sample studies document that children's abilities in experimentation, evidence evaluation, and causal inference show marked growth through early and middle childhood when instructional opportunities are provided (e.g., Zimmerman, 2007; Koerber et al., 2015). The pattern—particularly the difference between five-year-olds and older peers—parallels results demonstrating increasing competence in coordinating variables and interpreting evidence as schooling progresses and cognitive resources expand. These convergences suggest that the STS-AT is sensitive to developmental change, an important aspect of validity for instruments targeting emergent cognition.

At the subdomain level, children's relatively higher performance in analytical interpretation compared with metacognitive awareness resonates with work showing that explicit reflective monitoring lags behind more direct reasoning processes in early childhood. Observational and structured assessments indicate that while young children can engage in simple causal explanations, metacognitive monitoring and regulation are still consolidating during this period, often requiring scaffolds to surface reliably in assessment contexts. The STS-AT's metacognition items appear to detect this still-emerging capacity—consistent with developmental accounts of metacognition and with observational measurement traditions in early childhood.

Gender analyses yielded no statistically significant differences in total scores, a result aligned with multiple strands of recent evidence. Studies using comprehensive inventories of scientific reasoning in kindergarten and early primary school often report negligible or absent gender gaps in core reasoning competencies when opportunities to learn are comparable. A recent validation of the Science-K Inventory likewise reported no gender differences, and broader early-childhood work on foundational quantitative abilities similarly finds parity between girls and boys. Taken together, the lack of differences in our data supports the fairness of the STS-AT scores across genders in this age band and underscores the salience of equitable instructional experiences rather than presumed ability gaps.

Beyond psychometrics, the findings carry implications for curriculum and instruction. The four domains operationalized by the STS-AT—critical inquiry, hypothesis testing, analytical interpretation, and metacognitive awareness—map closely onto national policy frameworks that emphasize scientific practices, evidence use, and the cultivation of reflective learners from the earliest grades. Positioning assessment in service of instruction, educators can use STS-AT profiles to tailor inquiry experiences (e.g., structured prediction–verification tasks) and to explicitly scaffold metacognitive talk, thereby aligning classroom practice with contemporary standards for science learning.

Alongside its robust psychometric foundation, the STS-AT offers teachers concrete opportunities to apply its findings within classroom contexts. Teachers can employ the instrument not only as a diagnostic tool but also as a formative guide to support children's scientific learning. For instance, when a child demonstrates strength in hypothesis testing but relatively weaker metacognitive awareness, teachers may intentionally integrate reflective prompts such as "How did you decide that?" or "What might you do differently next time?" into everyday activities. Similarly, observations from the critical inquiry subscale can help teachers recognize children who are naturally curious questioners and design inquiry-based tasks that further cultivate this strength. By embedding the STS-AT into routine classroom interactions, educators can move beyond static assessment to foster individualized scaffolding, thereby aligning daily practices with curricular frameworks that emphasize inquiry and reflective thinking in early STEM education.

## Validity Considerations and Future Work

Although the present study provides multi-faceted validity evidence, several avenues can further strengthen the interpretive argument. First, longitudinal designs could establish sensitivity to growth at the individual level and permit the evaluation of predictive validity for later science achievement. Second, convergent and discriminant validity would benefit from multi-method batteries that include established early-years instruments (e.g., domain-specific reasoning tasks) and teacher reports to triangulate scores. Third, given the policy importance of

equity, future studies should evaluate measurement invariance explicitly across subgroups (e.g., gender, age bands, linguistic background) using multi-group CFA criteria recommended in the measurement literature (e.g., changes in CFI and RMSEA within recommended thresholds). Such work would extend the current fairness evidence and ensure that observed mean differences—when present—reflect true developmental or instructional effects rather than measurement artifacts.

### Limitations

The study's cross-sectional design restricts inferences about individual developmental trajectories; the school-based sampling frame within one national context may also limit generalizability across curricula and languages. While our reliability and factor structure are robust, future research should examine alternative models (e.g., bifactor or hierarchical structures) to test whether a general scientific thinking factor accounts for common variance alongside domain-specific factors, a question raised in large-sample studies of elementary-age learners. Incorporating response-process evidence (e.g., think-alouds) could further illuminate how young children interpret prompts, particularly in metacognitive items.

The STS-AT offers a psychometrically sound and instructionally meaningful assessment of early scientific thinking. By aligning with established developmental findings and contemporary standards while demonstrating strong reliability and construct validity, the instrument can support both diagnostic use in classrooms and research on the emergence of scientific reasoning. The absence of gender differences in this age range and the clear age-related progression suggests that high-quality, developmentally appropriate science experiences—especially those that explicitly elicit prediction, evidence coordination, and reflective talk—are likely to benefit all learners. Ongoing work on invariance, growth sensitivity, and cross-cultural applications will further consolidate the STS-AT's contribution to early STEM assessment and practice.

## Conclusion

This study developed and validated the Scientific Thinking Skills Assessment Tool (STS-AT), a multidimensional instrument designed to capture critical inquiry, hypothesis testing, analytical interpretation, and metacognitive awareness in children aged 5–8 years. Across a large and diverse Turkish sample, the STS-AT demonstrated strong psychometric properties, including internal consistency, inter-rater agreement, temporal stability, and a theoretically coherent four-factor structure confirmed through exploratory and confirmatory factor analyses. Together, these findings provide robust support for the instrument as a reliable and valid measure of early scientific thinking.

The STS-AT makes several contributions to early childhood science education and assessment. By integrating domains often measured separately, the tool allows for a more comprehensive evaluation of young children's reasoning skills. Its child-centered, play-based tasks and visual supports make it developmentally appropriate and accessible, while its standardized scoring procedures ensure reliable use across research and classroom contexts. For educators, the instrument provides a diagnostic framework that can guide the design of inquiry-based learning activities and targeted instructional interventions.

For policymakers and curriculum developers, it offers empirical evidence of the importance of fostering inquiry, reasoning, and reflection from the earliest years of formal schooling. Despite these strengths, limitations should be acknowledged. The sample was restricted to one national context, and cross-cultural validation will be essential to establish broader generalizability. Criterion validity was not assessed against external standardized measures, which should be addressed in future research. Longitudinal studies are also needed to examine the predictive validity of early scientific thinking for later STEM achievement and to evaluate growth trajectories at the individual level. Additionally, advanced psychometric approaches such as multi-group invariance testing and bifactor modeling would further illuminate the structure of scientific thinking across subgroups.

In sum, the STS-AT represents a timely and evidence-based contribution to early childhood research and practice. By capturing the complexity of young children's scientific reasoning, it fills a critical gap in existing assessment tools and offers a foundation for advancing theory, informing pedagogy, and promoting equitable opportunities for scientific learning. Ongoing refinement and cross-cultural application will ensure that the STS-AT continues to support the development of scientifically literate citizens from the earliest stages of education.

## Scientific Ethics Declaration

* The author declares that the scientific ethical and legal responsibility of this article published in EPESS journal belongs to the author.

* Ethical approval for the study was obtained from the Giresun University Social Sciences, Science, and Engineering Research Ethics Committee (Approval No: E-50288587-050.01.04-74549).

## Conflict of Interest

* The author declares that there is no conflict of interest

## Funding

* This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

## Acknowledgements and Notes

* This article was presented as an oral presentation at the International Conference on Research in Education and Technology ( www.icret.net ) held in Budapest/Hungary on August 28-31, 2025.

* The author gratefully acknowledges the children, families, and teachers who participated in this study. Special thanks are also extended to colleagues who provided valuable support during instrument refinement.

## References

AERA, APA, & NCME. (2014). *Standards for educational and psychological testing*. American Educational Research Association.

Bybee, R. W. (2013). *The case for STEM education: Challenges and opportunities*. NSTA Press.

Flavell, J. H. (1979). Metacognition and cognitive monitoring: A new area of cognitive–developmental inquiry. *American Psychologist, 34*(10), 906–911.

Gopnik, A., Meltzoff, A. N., & Kuhl, P. K. (2000). *The scientist in the crib: Minds, brains, and how children learn*. William Morrow.

Hu, L. T., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling, 6*(1), 1–55.

Koerber, S., Sodian, B., Thoermer, C., & Nett, U. (2015). Scientific reasoning in young children: Experimental knowledge and theory formation. *Developmental Psychology, 41*(3), 303–317.

Koerber, S., & Osterhaus, M. (2020). Developing a comprehensive assessment of scientific reasoning in preschool children: The Science-K Inventory. *Journal of Early Childhood Research, 18*(2), 123–138.

Köksal, Ö. (2022). Scientific thinking in young children: Development, culture, and education. In K. C. Trundle & M. Saçkes (Eds.), *Handbook of early childhood science education* (pp. 85–104). Routledge.

Kuhn, D. (2000). Metacognitive development. *Current Directions in Psychological Science, 9*(5), 178–181.

Kuhn, D. (2010). *The skills of argument*. Cambridge University Press.

Meisels, S. J., Liaw, F-R., Dorfman, A., & Nelson, R. (1995). The Work Sampling System: Reliability and validity of a performance assessment for young children. *Early Childhood Research Quarterly, 10*(3), 277–296.

National Research Council. (2012). *A framework for K-12 science education: Practices, crosscutting concepts, and core ideas*. National Academies Press.

NGSS Lead States. (2013). *Next generation science standards: For states, by states*. National Academies Press.

OECD. (2017). *OECD science, technology and innovation outlook 2016*. OECD Publishing.

Osterhaus, C., Koerber, S., & Sodian, B. (2023). The complex associations between scientific reasoning and advanced theory of mind: A developmental perspective. *Child Development, 94*(1), 113–130.

Osterhaus, C. (2023). Validating the Chinese version of the Science-K Inventory (SC-SKI): Factor structure, reliability, and measurement invariance. *Infant and Child Development, 32*(5), e2421.

Piaget, J. (1972). *The psychology of the child*. Basic Books.
Schraw, G., & Dennison, R. S. (1994). Assessing metacognitive awareness. *Contemporary Educational Psychology, 19*(4), 460–475.
Vygotsky, L. S. (1978). *Mind in society: The development of higher psychological processes*. Harvard University Press.
Whitebread, D., Coltman, P., Pasternak, D. P., Sangster, C., Grau, V., Bingham, S., Almeqdad, Q., & Demetriou, D. (2009). The development of two observational tools for assessing metacognition and self-regulated learning in young children. *Metacognition and Learning, 4*(1), 63–85
Zimmerman, C. (2007). The development of scientific thinking skills in elementary and middle school. *Developmental Review, 27*(2), 172–223.

## Author Information

**Elif Ozturk,**
Ph.D.
Indiana University of Bloomington, Visiting Scholar,
Department of Curriculum and Instruction, USA,
Contact e-mail:*elozturk@iu.edu*
Giresun University, Associate Professor, Department of Early Childhood Education,
Giresun, Türkiye,
Contact e-mail: *elif.ozturk@giresun.edu.tr*

**To cite this article:**

Ozturk, E. (2025). Assessing scientific thinking in early childhood: Development and validation of the scientific thinking skills assessment tool (STS-AT). *The Eurasia Proceedings of Educational and Social Sciences (EPESS), 45,* 88-97.

**Appendix A.**

Table Standardized factor loadings for the four-factor confirmatory factor analysis (N = 282)

| Item | Critical Inquiry | Hypothesis Testing | Analytical Interpretation | Metacognitive Awareness |
|---|---|---|---|---|
| ESM1.62 | — | — | — | |
| ESM2.65 | — | — | — | |
| ESM3.71 | — | — | — | |
| H1 | — | .59 | — | — |
| H2 | — | .63 | — | — |
| H3 | — | .67 | — | — |
| AY1 | — | — | .74 | — |
| AY2 | — | — | .79 | — |
| AY3 | — | — | .72 | — |
| MF1 | — | — | — | .68 |
| MF2 | — | — | — | .71 |
| MF3 | — | — | — | .69 |

*Note.* All factor loadings are standardized estimates and statistically significant ($p < .001$). Dashes (—) indicate non-specified loadings in the four-factor model. Model fit indices: $\chi^2/df = 1.92$, RMSEA = .04, CFI = .95, TLI = .93, SRMR = .06.