Journal of Social and Scientific Education, Vol. 3 No. 3 (2026) : 827-842 ISSN 3047-0684
Available online at: https://doi.org/10.58230/josse.v3i3.1107

Invisible Reasoning: What Early Childhood Teachers Notice, and Overlook, in Children'S Scientific Thinking

Subhan
Universitas Islam Negeri Palopo, Indonesia
subhan@uinpalopo.ac.id

ABSTRACT

Early childhood teachers are increasingly asked to assess and support children's scientific thinking, yet little is known about what teachers actually attend to when they describe a child's thinking in their own words. This study examined how teachers' spontaneous narrative notes about preschool children correspond to their structured ratings of scientific thinking, with particular attention to hypothesis and experimentation skills. A convergent mixed-methods design was used. Teachers from 277 early childhood institutions in Indonesia rated 295 children aged 4–6 years on a 24-item instrument covering attention, working memory, problem solving, observation, and hypothesis and experimentation, and 225 child-specific open-ended notes were analysed with directed content analysis; the two strands were integrated through cross-tabulation, correlation, and regression. Hypothesis and experimentation was the lowest-rated dimension (M = 3.31), yet it was almost absent from teachers' narratives: no note mentioned prediction, hypothesising, or revising an idea, and only 11.1% referred to any scientific reasoning process. Notes were dominated by global evaluations (43.6%) and attention or behavioural-regulation concerns (22.7%). For children in the lowest quartile of rated hypothesis and experimentation, teachers wrote about attention or regulation problems in 30.6% of notes but about reasoning in only 4.8%; in the highest quartile, reasoning was mentioned in 23.9%. Attention concerns did not predict hypothesis scores once rated attention was controlled, whereas mentions of reasoning added unique information beyond all four other rated dimensions. The findings suggest that children's emerging scientific reasoning is largely invisible in teachers' everyday descriptions, and that difficulties in this domain tend to be described as problems of focus, although the cross-sectional, correlational design cannot establish whether this reflects genuine attentional constraints, teachers' perception, or both. Professional development should therefore train teachers to notice and document children's predictions, explanations, and revisions explicitly.

Keywords: Scientific Thinking, Teacher Noticing, Early Childhood Assessment, Hypothesis Testing, Mixed Methods

Corresponding Author: [1]Subhan - [1]subhan@uinpalopo.ac.id

How to cite : Subhan, S. (2026). Invisible Reasoning: What Early Childhood Teachers Notice, and Overlook, in Children's Scientific Thinking. *Journal of Social and Scientific Education, 3*(3), 827-842. https://doi.org/10.58230/josse.v3i3.1107

INTRODUCTION

Scientific thinking, understood as the intentional coordination of ideas and evidence, is no longer regarded as a late achievement of formal schooling. Over the past three decades developmental research has shown that preschool children generate and test hypotheses, distinguish confounded from unconfounded evidence, and revise their explanations when outcomes are inconsistent with their expectations (Gopnik, 2012; Legare, 2012; Schulz & Bonawitz, 2007). Kindergarten children can evaluate evidence and generate hypotheses in simple domain-general tasks (Piekny & Maehler, 2013), and with appropriate support can learn to construct unconfounded experiments (van der Graaf et al., 2015). Spontaneous exploratory play has been described as a form of intuitive experimentation in which children

This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.
*Copyright (c) 2026 **Subhan**

select actions that are informative about causal structure (Cook et al., 2011). At the same time, these skills develop slowly and unevenly across childhood, and the coordination of hypothesis and evidence remains difficult well into the elementary years (Koerber et al., 2015; Zimmerman, 2007).

This developmental picture has made early science education a policy priority. Children's early science experiences are thought to lay foundations for later scientific literacy (Greenfield et al., 2009; Larimore, 2020), and domain-general cognitive resources such as executive function and causal reasoning appear to support preschoolers' emerging scientific competence (Bauer & Booth, 2019). Yet a consistent finding in the literature is that what happens in classrooms lags behind what children are capable of. Science is taught less frequently than literacy or mathematics in preschool classrooms, and teachers report lower self-efficacy for science than for literacy (Gerde et al., 2018). Pre-kindergarten teachers value science for young children but often feel uncomfortable teaching it (Maier et al., 2013; Pendergast et al., 2017), and pre-school teachers hold diverse and sometimes vague views about what a curious child looks like and how curiosity can be fostered (Spektor-Levy et al., 2013). Even when science activities are offered, their mere frequency is a weak predictor of children's later science achievement (Saçkes et al., 2011), which suggests that the quality of teachers' engagement with children's thinking matters more than the quantity of science time.

A central component of that quality is what teachers notice. Research on professional noticing defines it as a set of interrelated skills: attending to the details of children's strategies, interpreting the understandings those strategies reflect, and deciding how to respond on that basis (Jacobs et al., 2010). Noticing has been studied extensively in mathematics, but much less in early science. This gap matters because scientific reasoning in young children is often fleeting and embedded in play and talk, a prediction voiced in passing or a change of mind after an unexpected result (Gerde et al., 2013; Legare et al., 2016). If teachers do not attend to such moments, they cannot build on them. Classroom studies indeed show that preschool teachers both promote and suppress children's curiosity during math and science activities, often without recognising the epistemic opportunities in front of them (Evans et al., 2023).

A second, related body of research concerns the accuracy and structure of teacher judgments. Teachers judge academic achievement with reasonable accuracy, with a mean correlation of about .63 between judgments and test scores (Südkamp et al., 2012), but they are considerably less accurate when judging cognitive abilities (Machts et al., 2016). Teacher judgments are also prone to cross-contamination: students' engagement influences judgments of their achievement and vice versa (Kaiser et al., 2013), and salient disruptive behaviour inflates ratings of other problem behaviours, a well-documented halo effect (Abikoff et al., 1993; Hartung et al., 2006). Behaviour ratings and performance-based measures of cognitive functions often capture different aspects of the child (Toplak et al., 2013). Taken together, these findings suggest that the most visible features of a child's classroom behaviour, especially attention and conduct, may shape how teachers perceive less visible cognitive processes.

Most studies of early scientific thinking rely on either direct child assessment or structured teacher ratings. Structured ratings show how teachers score children when prompted with a specific item, but not what teachers spontaneously attend to when they describe a child in their own words. Narrative records of that kind, often called anecdotal notes, are a routine part of formative assessment in early childhood settings, including in Indonesia,

and they are the material from which teachers plan the next learning experience. Curiosity and inquiry in young children are themselves difficult to operationalise (Jirout & Klahr, 2012), and curiosity is related to some inquiry processes but not others (van Schijndel et al., 2018). It is therefore an open question whether teachers' narratives capture the specific reasoning processes, such as predicting, explaining, and revising, that distinguish scientific thinking from general engagement.

The present study addresses this gap. It builds on the author's earlier work with the Scientific Thinking Assessment for Early Childhood (STAEC), a teacher-rated instrument that, in a sample of 105 preschoolers in South Sulawesi, showed general cognitive capacities (attention, working memory, problem solving) rated higher than domain-specific science competencies, with working memory most closely linked to prediction, reasoning, and experimentation (Subhan et al., 2026). That study, like most work in the field, relied on structured ratings only, and its authors noted that teacher ratings may introduce bias and cannot capture children's reasoning as it unfolds. The present study draws on a larger, nationally distributed dataset in which teachers rated children's scientific thinking with the STAEC and also wrote an optional open-ended note about each child's behaviour or abilities. The same data collection also supports a companion manuscript on the structural relations among the five rated dimensions; how the two papers differ in data, variables, and claims is set out in the Data Overlap Statement in the Methods. The present study does not examine that structure; instead, it turns to the teachers themselves and asks what the weakness in hypothesis and experimentation looks like from the teacher's perspective. Studies that relate teachers' free narratives to their own structured ratings of the same children are rare in any domain or age group. A recent review of 35 studies on preschool teachers' noticing found that the field has concentrated on mathematics and has relied mainly on video vignettes of other teachers' classrooms, so that noticing within teachers' own everyday interactions and records remains largely unexamined (Stiewe & Ehm, 2026). Outside education, a recent clinical study combined open-ended teacher narratives with a standardised teacher rating scale and found that the two sources carried largely independent information about children's attention problems (Karacan et al., 2026). The present study therefore does not claim to introduce the integration of narratives and ratings as such. Its contribution lies in applying that integration to a content domain and a type of record where it has not been used, namely the routine notes that early childhood teachers write about their own pupils, and in asking whether these notes capture the specific reasoning processes that teachers rate as weakest. This makes it possible to examine not only how teachers rate scientific reasoning when prompted but whether they see it unprompted. Three research questions guided the study: (1) What aspects of children's functioning do teachers spontaneously describe in their narrative notes, and how often do they refer to scientific reasoning processes? (2) How does the content of teachers' narratives vary with children's rated hypothesis and experimentation skills? (3) Do narrative references to attention problems and to scientific reasoning carry information about rated hypothesis and experimentation beyond the structured ratings of other cognitive dimensions?

METHODS OF RESEARCH

Research Design

The study used a convergent mixed-methods design in which quantitative and qualitative data were collected in the same instrument and integrated at the analysis and interpretation stages (Fetters et al., 2013). The quantitative strand consisted of structured teacher ratings of children's scientific thinking; the qualitative strand consisted of teachers' open-ended notes about the same children. Integration was achieved by transforming the coded narratives into category indicators and relating them to the rating data through cross-tabulation, correlation, and regression, and by interpreting the resulting patterns with illustrative excerpts. This design was chosen because the research questions concern the correspondence between what teachers rate when prompted and what they describe spontaneously, which neither strand could answer alone.

Participants

Data were collected between December 2025 and January 2026 through an online form distributed to early childhood teachers across Indonesia. Of 300 submissions, one was excluded because consent was declined and four because ten or more of the 25 ratings were missing, leaving 295 children aged 4 to 6 years (4 years, n = 85; 5 years, n = 134; 6 years, n = 75; age not reported, n = 1), 149 girls and 146 boys. The children were rated by teachers from 277 early childhood institutions, including kindergartens (TK), Islamic kindergartens (RA), playgroups (KB), and integrated PAUD units, so that the data are close to one child per teacher. Teachers had known the child for more than six months (n = 130), about six months (n = 104), three to six months (n = 46), or less than three months (n = 14). Of the 295 children, 248 had a narrative note, and 225 of these notes were child-specific (see Data Analysis). The 70 children without a usable note (47 with no note and 23 with a note that was not child-specific) did not differ from the 225 children with a child-specific note on any rated dimension (Welch's t-tests, all $p \geq .075$, $|d| \leq 0.25$), and their mean hypothesis and experimentation rating was virtually identical ($M = 3.35$ vs. 3.30, $p = .665$). The availability of usable notes therefore does not appear to be selective with respect to the ratings.

Instruments

Children's scientific thinking was assessed with the Scientific Thinking Assessment for Early Childhood (STAEC), a 25-item teacher-rated instrument developed through expert review and piloting with 30 teachers (Subhan et al., 2026). Items were rated on a five-point frequency scale (1 = never to 5 = always) and grouped into five dimensions: attention (items 1–5), working memory (items 6–10), problem solving and reasoning (items 11–15), observation (items 16–18), and hypothesis and experimentation (items 19–25). This five-dimension scoring combines the prediction-and-reasoning and experimentation subscales of the original six-dimensional model into a single hypothesis and experimentation dimension, because both concern the coordination of ideas and evidence that is the focus of the present study. The hypothesis and experimentation dimension comprises making predictions, giving reasons for predictions, considering other possible outcomes, trying different ways to obtain different results, showing interest in simple experiments, explaining observations, and revising opinions after seeing a different result.

Item 5 ("the child is easily distracted by other sounds or movements") was the only negatively worded item. Preliminary analyses showed that its raw score correlated positively, not negatively, with item 1 ("the child is able to sustain attention"), r = .18, and with the remaining attention items, r = .16, p = .006, and that reverse-scoring it reduced the internal consistency of the attention dimension to α = .69. Because these patterns indicate that the item was not answered as a reverse-worded item, it was excluded, and attention was scored from items 1–4. Because this decision rests on patterns in the present sample, all integrated analyses were repeated with attention scored from all five items, once with item 5 reverse-scored and once with its raw score. The coefficients for attention concerns and for scientific reasoning references changed by no more than 0.04, and scientific reasoning references still predicted hypothesis and experimentation over and above all four other rated dimensions (b = 0.27, SE = 0.13, p = .040 in both versions). Item-level data from the earlier validation sample of 105 preschoolers were not available for re-analysis, so it could not be checked whether the item behaved in the same way there. Internal consistency in the analytic sample was good to excellent for all dimensions: attention α = .84, working memory α = .86, problem solving α = .87, observation α = .85, and hypothesis and experimentation α = .91. Dimension scores were computed as item means; 20 of 7,080 item responses (0.3%) were missing and were handled by averaging the available items.

The qualitative data came from an optional open-ended item placed after the rating scale: "Additional notes on the child's behaviour or abilities." The item did not mention science or scientific thinking, so the responses reflect what teachers chose to write without a content prompt.

Data Analysis

The narrative notes were analysed with directed content analysis (Hsieh & Shannon, 2005). Each note was first screened for whether it described the individual child, and 23 notes that contained only general statements about children, parenting, or the teacher's own practice, or that were unrelated to the child, were set aside, leaving 225 child-specific notes. A codebook of nine categories was then applied, namely attention and behavioural-regulation concerns (C1), positive attention or persistence (C2), socio-emotional functioning (C3), task compliance and independence (C4), academic skills, memory, and language (C5), curiosity and interest (C6), scientific reasoning processes such as observing, explaining, reasoning about cause and effect, problem-solving strategies, or experimenting (C7), global evaluative statements (C8), and activity level (C9). Each note could receive several codes. Coding was carried out by the author with keyword-assisted screening followed by a full manual reading of every note. To check the consistency of the two codes on which the main conclusions rest, a random sample of 40 child-specific notes (10 from each quartile of rated hypothesis and experimentation, presented in random order with names removed) was coded a second time, independently and without access to the rating data or to the original codes. Because no trained second human coder was available, this second coding was carried out by a large language model (Claude, Anthropic) that received only the codebook and the note texts. Agreement with the original codes was 95.0% (Cohen's κ = .80) for attention and behavioural-regulation concerns (C1) and 100% (κ = 1.00) for scientific reasoning processes (C7), although the latter estimate rests on only six positive cases in the subsample. To strengthen the dependability of

the coding, the codebook was applied in two passes: keyword screening first flagged candidate notes for each category, and every note was then read in full to confirm, remove, or add codes, with particular attention to negations (for example, distinguishing "focused" from "not yet focused") and to notes that described the teacher's own practice rather than the child. A coding sheet listing each note, its codes, and the corresponding rating data was retained as an audit trail. In addition, a lexical search was conducted for terms that denote hypothesis and experimentation processes most directly (prediction, guessing, hypothesis, revising, changing one's opinion, cause, experiment).

For the quantitative strand, descriptive statistics and paired t-tests compared hypothesis and experimentation with the other dimensions. For integration, children with child-specific notes were divided into quartiles of rated hypothesis and experimentation, and the proportion of notes in each category was compared across quartiles with chi-square tests and Kendall's τ for monotonic trend. Because some expected cell frequencies were small (the smallest was 4.7, for activity level in the highest quartile, and the smallest for scientific reasoning was 5.1), every chi-square result was also checked with the Freeman-Halton extension of Fisher's exact test. Point-biserial correlations related each category to the five rated dimensions. Finally, ordinary least squares regressions with standardised hypothesis and experimentation scores as the outcome tested whether the presence of attention concerns (C1) or of scientific reasoning references (C7) in a note explained variance beyond rated attention alone and beyond all four other rated dimensions. To rule out verbosity as an explanation, these models were re-estimated with the log-transformed word count of each note as an additional covariate, and multicollinearity and the robustness of the coefficients were examined with variance inflation factors and heteroscedasticity-consistent (HC3) standard errors. As a sensitivity check, analyses were repeated after excluding ratings by teachers who had known the child for less than three months and one response labelled by its author as a coursework submission. Analyses were conducted in Python (pandas, SciPy).

Ethical Considerations

The form opened with an informed-consent statement explaining the purpose of the study, the confidentiality of responses, and the voluntary nature of participation. Only responses from teachers who gave consent were analysed. Children's names were removed before analysis, and excerpts are reported without identifying information.

Data Overlap Statement

This manuscript and a companion manuscript, "The Cognitive Architecture of Scientific Thinking in Early Childhood: A Teacher-Rated Assessment of Life Science Reasoning," draw on the same data collection, namely the STAEC ratings of 295 children submitted through a single online form. The two papers differ in their primary data, analyses, and claims. The companion manuscript uses only the rated items; it examines the internal structure of the instrument, the relations among the rated dimensions, including mediation among them, and differences by age and gender, and it does not analyse the open-ended notes. The present manuscript takes the 225 child-specific narrative notes as its primary data and uses the rated dimensions only as descriptive context (Table 1) and as covariates and criterion against which the coded narratives are compared. The material common to both papers is limited to the sample

description, the descriptive statistics of the rated dimensions, and the baseline regression of hypothesis and experimentation on the four other dimensions, which is reported here only as the control model for testing the narrative indicators. No structural, mediation, or group-difference result from the companion manuscript is reported or relied upon in this paper. The overlap is disclosed here so that editors and readers can judge the independence of the two contributions.

RESULTS AND DISCUSSION

Hypothesis and Experimentation as the Weakest Rated Dimension

Table 1 presents the descriptive statistics for the five rated dimensions. Hypothesis and experimentation received the lowest mean rating (M = 3.31, SD = 0.81) and was rated significantly lower than working memory, $t(294) = 11.67$, $p < .001$, $dz = 0.68$, observation, $t(294) = 10.48$, $p < .001$, $dz = 0.61$, attention, $t(294) = 6.83$, $p < .001$, $dz = 0.40$, and problem solving, $t(294) = 5.78$, $p < .001$, $dz = 0.34$. At item level, the lowest-rated behaviours were considering other possible outcomes (M = 3.03; 8.1% of children rated 1), making predictions (M = 3.18), giving reasons for predictions (M = 3.20), and revising an opinion after seeing a different result (M = 3.26). This pattern, in which hypothesis generation and belief revision lag behind observation and memory, is consistent with developmental evidence that the coordination of hypothesis and evidence is the slowest-developing component of scientific thinking (Koerber et al., 2015; Piekny & Maehler, 2013; Zimmerman, 2007).

Table 1 Descriptive Statistics and Reliability of the Rated Dimensions (N = 295)

Dimension	Items	α	M	SD
Attention	1–4	.84	3.62	0.81
Working memory	6–10	.86	3.74	0.76
Problem solving and reasoning	11–15	.87	3.51	0.78
Observation	16–18	.85	3.66	0.81
Hypothesis and experimentation	19–25	.91	3.31	0.81

Note. Item 5 was excluded (see Instruments). Ratings on a 1–5 scale. Correlations among the five rated dimensions ranged from .53 (attention with hypothesis and experimentation) to .80 (working memory with problem solving), all $p < .001$.

What Teachers Write: Global Evaluations and Attention, Rarely Reasoning

Teachers' child-specific notes were short (median 10 words) and carried on average 1.6 codes. Note length was unrelated to every rated dimension (Spearman's ρ between −.04 and .07, all $p > .29$), including hypothesis and experimentation ($\rho = -.01$, $p = .899$). Notes that referred to scientific reasoning were, however, longer than other notes (median 20 vs. 10 words), so length was controlled in the regression analyses reported below. Table 2 summarises the categories. The most frequent category was global evaluation (43.6%), statements such as "already developing well according to age" or "still needs guidance" that judge the child without describing what the child does. The second most frequent was attention and behavioural-regulation concerns (22.7%), for example "the child is often not focused during activities in class" or "his concentration is easily broken by his friends." Task compliance and independence (18.2%), academic skills and memory (16.4%), curiosity and interest (16.4%), and socio-emotional functioning (16.0%) were mentioned at similar rates.

Table 2 Categories of Teachers' Child-Specific Narrative Notes (N = 225)

Category	n	%	Illustrative excerpt (translated)
C8 Global evaluation	98	43.6	"Already developing very well according to age."
C1 Attention and behavioural-regulation concerns	51	22.7	"His concentration is easily broken by his friends."
C4 Task compliance and independence	41	18.2	"Always does the tasks given by the teacher."
C5 Academic skills, memory, and language	37	16.4	"Can already spell the letters from A to Z."
C6 Curiosity and interest	37	16.4	"Shows great interest and often asks how things work."
C3 Socio-emotional functioning	36	16.0	"Lacks confidence and is moody."
C7 Scientific reasoning processes	25	11.1	"Diligently observes and can explain the results clearly."
C9 Activity level	23	10.2	"Very active in class."
C2 Positive attention or persistence	11	4.9	"Careful and focused in the learning process."

Note. Categories are non-exclusive; a note could receive several codes. Seven notes received no code.

Scientific reasoning processes were mentioned in only 25 notes (11.1%), and most of these referred to observing or explaining ("she diligently observes and can explain the results clearly"). The lexical search sharpened this picture. Across all 225 child-specific notes, no teacher used any word for predicting, guessing, hypothesising, or revising an idea. Hypothetical reasoning appeared once, in a note describing a child who "often wonders that if an animal is like this, then it will be like that"; cause and effect appeared twice and experimentation three times. In other words, the very behaviours that make up the weakest rated dimension, predicting, justifying predictions, considering alternatives, and revising beliefs, were essentially absent from teachers' spontaneous descriptions of children.

This finding extends earlier work showing that early childhood teachers feel less prepared for science than for literacy (Gerde et al., 2018; Maier et al., 2013). When teachers write about a child freely, their descriptive vocabulary is organised around behaviour, compliance, and general progress, not around the epistemic moves of scientific inquiry. In the terms of the professional-noticing framework, the first component, attending to the details of children's thinking, is already constrained at the level of what teachers record (Jacobs et al., 2010). The relatively frequent references to curiosity (16.4%) suggest that teachers do recognise children's interest and questioning; what is missing is the next step from noticing that a child is curious to noticing how the child reasons.

Attention as the Dominant Lens for Children with Weak Hypothesis Skills

Table 3 and Figure 1 show how the content of teachers' notes varied across quartiles of rated hypothesis and experimentation. Two categories showed significant monotonic trends. References to attention and behavioural-regulation concerns were most frequent for children in the two lower quartiles (30.6% and 35.0%) and dropped to about 11% in the upper quartiles, $\chi^2(3) = 15.90$, $p = .001$, exact $p = .001$, $\tau = -.20$, $p < .001$. References to scientific reasoning showed the opposite gradient, rising from 4.8% in the lowest quartile to 23.9% in the highest, $\chi^2(3) = 11.38$, $p = .010$, exact $p = .015$, $\tau = .19$, $p = .002$. Curiosity showed a similar but non-significant rise (11.3% to 23.9%; $\tau = .11$, $p = .082$), and none of the other categories varied significantly across the four quartiles. Global evaluation, however, showed a non-

monotonic pattern that a trend test can easily miss. It was fairly flat across the three lower quartiles (41.9%, 38.3%, and 38.6%) and then rose sharply to 58.7% in the highest quartile. A post hoc comparison of the highest quartile with the three lower quartiles combined (39.7%) was significant, $\chi^2(1) = 5.39$, $p = .020$, Fisher's exact $p = .030$, odds ratio = 2.16, although this contrast was suggested by the data and should be read as exploratory. The jump suggests that strong reasoners attract more generic praise as well as more specific descriptions of reasoning, so that a favourable overall impression may lead teachers to add summary approval when they write about the children they regard most highly.

Table 3 Percentage of Notes in Each Category by Quartile of Rated Hypothesis and Experimentation

Category	Q1 (n = 62)	Q2 (n = 60)	Q3 (n = 57)	Q4 (n = 46)	$\chi^2(3)$	τ
C1 Attention/regulation concerns	30.6	35.0	10.5	10.9	15.90**	−.20***
C7 Scientific reasoning	4.8	6.7	12.3	23.9	11.38*	.19**
C6 Curiosity and interest	11.3	15.0	17.5	23.9	3.21	.11
C8 Global evaluation	41.9	38.3	38.6	58.7	5.59	.09
C4 Task compliance	24.2	18.3	17.5	10.9	3.17	−.10
C5 Academic, memory, language	16.1	13.3	19.3	17.4	0.80	.03
C3 Socio-emotional	16.1	16.7	21.1	8.7	2.93	−.04
C9 Activity level	14.5	11.7	8.8	4.3	3.24	−.11

*Note. Quartile ranges of rated hypothesis and experimentation: Q1 = 1.00–2.86, Q2 = 3.00–3.43, Q3 = 3.57–3.86, Q4 = 4.00–5.00. τ = Kendall's tau for monotonic trend. C2 is omitted (n = 11). Exact Freeman-Halton p values were .001 for C1, .015 for C7, and between .14 and .85 for all other categories. *p < .05. **p < .01. ***p < .001.*

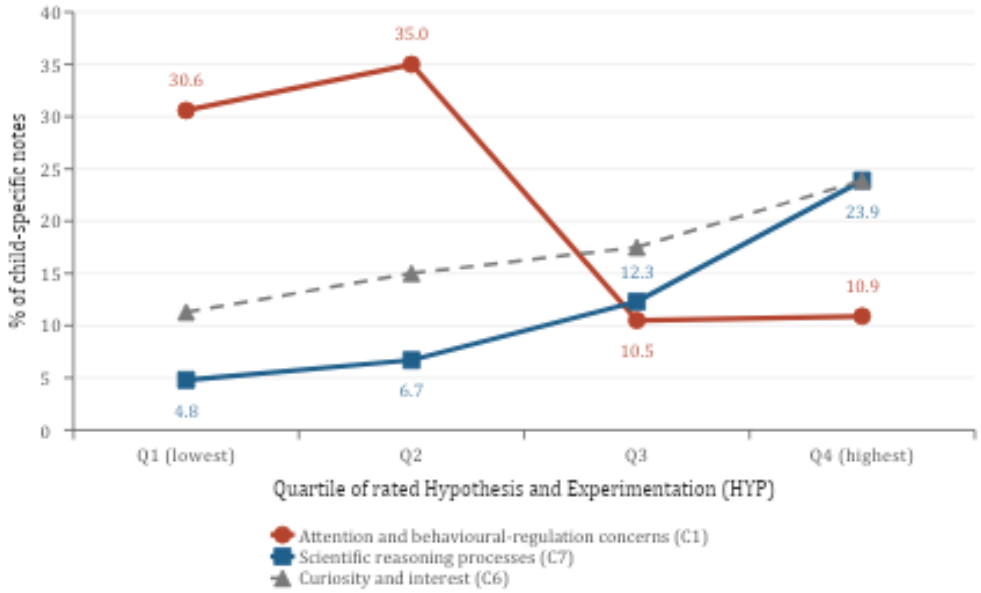


Figure 1 Percentage of Child-Specific Notes Referring to Attention Concerns, Scientific Reasoning, and Curiosity Across Quartiles of Rated Hypothesis and Experimentation

The asymmetry is the core result of this study. When a child is a strong scientific reasoner, the teacher's narrative tends to describe the reasoning: "the child is able to analyse and explain what has been seen or observed," or "the child's reasoning is developing very well, as seen in how the child explains the cause and effect of a phenomenon in simple language." When a child is weak in hypothesis and experimentation, the narrative rarely describes the

reasoning difficulty itself. Of the 62 notes about children in the lowest quartile, 19 described attention or regulation problems, whereas only three referred to any reasoning process, and only one of these described a difficulty in reasoning ("the child is not yet consistent in understanding cause–effect relationships and needs repeated stimulation"). Typical notes in this quartile read "the child lacks concentration or focus during activities," "active and gets bored quickly," or "still needs guidance." In other words, children's weak hypothesis skills tend to be visible to teachers as problems of focus and behaviour rather than as problems of reasoning.

Two interpretations of this pattern are possible, and they are not mutually exclusive. The first is substantive: attention and executive resources do support early scientific reasoning (Bauer & Booth, 2019), so children with attention difficulties may genuinely have more trouble predicting and testing ideas. The second concerns teacher perception: disruptive or inattentive behaviour is highly salient and is known to colour teachers' judgments of other characteristics (Abikoff et al., 1993; Kaiser et al., 2013). The regression analyses, reported next, help to separate these accounts.

Do Narratives Carry Information Beyond the Ratings?

Children whose notes mentioned attention concerns were rated lower on all five dimensions, with effects of similar size across dimensions (point-biserial r from −.21 for hypothesis and experimentation to −.28 for attention; Cohen's d from −0.49 to −0.67; all $p < .01$). This uniform pattern suggests that a teacher's concern about attention is associated with a lower overall impression of the child rather than with a specific cognitive weakness. Consistent with this, attention concerns did not predict hypothesis and experimentation once rated attention was controlled, $b = -0.13$, $SE = 0.14$, $p = .361$, nor once all four other rated dimensions were controlled, $b = -0.04$, $SE = 0.10$, $p = .689$.

References to scientific reasoning behaved differently. They were specifically associated with rated hypothesis and experimentation ($r = .20$, $p = .002$) and observation ($r = .17$, $p = .013$), but not with attention ($r = .06$), working memory ($r = .12$), or problem solving ($r = .12$). They predicted hypothesis and experimentation over and above rated attention, $b = 0.55$, $SE = 0.17$, $p = .002$, and, more stringently, over and above all four other rated dimensions together, $b = 0.27$, $SE = 0.13$, $p = .041$ (model $R^2 = .64$). Because the outcome and the four rated predictors come from the same teacher and the same instrument, much of this R^2 reflects shared rater variance; the rated dimensions correlated between .53 and .80 (see note to Table 1), and the variance inflation factors were 2.35 to 3.65 for the rated dimensions and 1.04 for the scientific reasoning indicator. Shared rater variance inflates what the rated predictors explain, but it works against the narrative indicator, which comes from a different source and was almost independent of the other predictors. Its unique contribution was accordingly small ($\Delta R^2 = .007$, partial $r = .14$), yet it was stable. The coefficient remained significant with HC3 standard errors ($b = 0.27$, $SE = 0.14$, $p = .049$), when the four rated dimensions were replaced by their mean ($b = 0.34$, $SE = 0.13$, $p = .013$), and when note length was added as a covariate, both with rated attention alone ($b = 0.59$, $SE = 0.18$, $p = .001$) and with all four rated dimensions ($b = 0.32$, $SE = 0.14$, $p = .018$). Note length itself did not predict hypothesis and experimentation in either model ($p \geq .143$), and controlling it did not alter the null result for attention concerns. In other words, when a teacher spontaneously writes about a child's reasoning, that

observation contains information about the child's hypothesis and experimentation skills that is not already captured by the child's ratings on the other cognitive dimensions. Teacher narratives are therefore not merely redundant with ratings; when they do describe reasoning, they are diagnostically informative. The problem is that they do so rarely.

The sensitivity analysis supported these conclusions. Excluding ratings by teachers who had known the child for less than three months and one coursework-labelled response (remaining n = 215), the gradients were virtually unchanged: attention concerns 32.8%, 34.5%, 11.3%, and 10.9% across quartiles, $\chi^2(3) = 15.22$, $p = .002$; scientific reasoning 5.2%, 6.9%, 13.2%, and 23.9%, $\chi^2(3) = 10.50$, $p = .015$.

Theoretical and Practical Implications

Theoretically, the findings bring together two literatures that have developed largely separately: developmental research on early scientific reasoning and educational research on teacher noticing and judgment. Developmental studies show that preschoolers can predict, explain inconsistent evidence, and revise their beliefs (Legare, 2012; Legare et al., 2016; Schulz & Bonawitz, 2007). The present data indicate that, in everyday classroom documentation, these capacities are seldom named. The weakness in hypothesis and experimentation found in the present ratings, in the earlier STAEC study (Subhan et al., 2026), and in previous developmental work may therefore have a pedagogical as well as a developmental component: skills that are not noticed are unlikely to be scaffolded. This interpretation is consistent with evidence that the quality of teacher–child interaction around science, rather than the frequency of science activities, is what matters for children's learning (Saçkes et al., 2011), and that teachers' responses can either promote or suppress children's curiosity (Evans et al., 2023).

The results also add a specific form to the halo literature. Previous research has shown that disruptive behaviour inflates ratings of other problem behaviours (Abikoff et al., 1993; Hartung et al., 2006) and that engagement colours achievement judgments (Kaiser et al., 2013). Here, attention concerns were associated with lower ratings across all cognitive dimensions to a similar degree and, in narrative form, appeared to stand in for descriptions of reasoning. For the children who most need support in hypothesis and experimentation, the teacher's account of the problem is therefore may tend to point toward behaviour management rather than toward cognitive scaffolding.

Practically, three implications follow. First, professional development in early science should include explicit training in noticing scientific reasoning, for example by showing teachers short episodes of play and asking them to identify predictions, explanations, and changes of mind, as has been done successfully for children's mathematical thinking (Jacobs et al., 2010). Second, anecdotal records and observation formats in early childhood settings could include prompts that direct attention to reasoning ("What did the child predict? What did the child do when the result was unexpected?"), so that documentation captures processes and not only general progress. Third, when a teacher identifies a child as "not focused," this may be a useful moment to check whether the child is also being given opportunities to predict and test ideas, using the playful and question-based approaches that are known to elicit scientific reasoning in young children (Gerde et al., 2013; Larimore, 2020).

Limitations and Future Research

Several limitations should be noted. First, all quantitative data were teacher ratings, a limitation already noted for the STAEC (Subhan et al., 2026), so the association between narratives and ratings partly reflects a single informant's perspective; direct child assessments of scientific reasoning, such as dynamic tasks for the control-of-variables strategy (van der Graaf et al., 2015), would allow a stronger test of whether attention-focused narratives reflect perception or genuine differences. Second, the notes were optional and short, and the open-ended prompt asked about "behaviour or abilities," which may have encouraged behavioural descriptions; a prompt that explicitly invited descriptions of thinking might yield different content, although the fact that a quarter of notes about the strongest reasoners did describe reasoning suggests that the prompt did not preclude it. Third, the sample was a convenience sample of teachers who responded to an online form, with almost one child per teacher, so teacher-level characteristics such as training and experience could not be modelled. Fourth, the full set of notes was coded by a single researcher, and the reliability check relied on an AI-assisted second coding of a subsample rather than on a trained human coder; future studies should double-code the complete material with human coders. In addition, the categories were assigned at note level rather than at the level of individual utterances. Future research should combine direct child assessment, observation of teacher–child interaction during science activities, and structured noticing tasks, and should test whether noticing-focused professional development changes both teachers' documentation and children's hypothesis and experimentation skills.

CONCLUSIONS

This study examined whether early childhood teachers' spontaneous narratives about children capture the scientific reasoning that their structured ratings identify as children's weakest area. Integrating ratings of 295 children with 225 child-specific teacher notes, it found that hypothesis and experimentation was both the lowest-rated dimension and the least visible one in teachers' narratives: no teacher wrote about prediction, hypothesising, or revising ideas, and references to any reasoning process were rare. For children with weak hypothesis skills, teachers wrote mainly about attention and behaviour, whereas for strong reasoners they described the reasoning itself, and these descriptions carried information beyond the ratings of all other cognitive dimensions. The study thus contributes evidence that children's scientific reasoning is largely invisible in everyday teacher documentation and that difficulties in this domain tend to be described as problems of focus. Because the design was cross-sectional and correlational, this pattern may reflect genuine attentional constraints on children's reasoning, a tendency in how teachers perceive and describe children, or both, and the present data cannot separate these accounts. The conclusions are limited by the reliance on a single informant, the brevity of optional notes, and a convenience sample with nearly one child per teacher. Future research should combine direct child assessment with observations of teacher–child interaction and test noticing-focused professional development that trains teachers to recognise and document children's predictions, explanations, and revisions.

ACKNOWLEDGMENT

The author would like to express sincere gratitude to the kindergarten teachers in South Sulawesi and in other regions of Indonesia who voluntarily took the time to complete the questionnaire and to share their observations of the children in their care. This study would not have been possible without their participation.

AI ACKNOWLEDGMENT

The author declares that he used Claude (Anthropic) for assistance with data analysis, blind second coding of the reliability subsample, drafting, and English-language editing of this manuscript. Despite using AI, the author affirms that he is the sole author of this article and is fully responsible for its content, as outlined in the COPE recommendations.

PARTICIPANT CONSENT

This research was conducted based on informed consent from all participants. Participants received an explanation of the purpose of the study, the confidentiality of their responses, and their right to withdraw at any time without consequence before completing the questionnaire. Only responses from teachers who gave consent were analysed, and no identifying information about the children is reported.

CONFLICT OF INTEREST

The author declares no conflict of interest.

AUTHOR CONTRIBUTIONS

S. conceptualized the research idea, developed the instrument, collected the data, conducted the quantitative and qualitative analyses, and wrote, revised, and approved the final version of this manuscript. The percentage contribution for the conceptualization, drafting, and revision of this paper is as follows: S.: 100%.

REFERENCES

Abikoff, H., Courtney, M., Pelham, W. E., & Koplewicz, H. S. (1993). Teachers' ratings of disruptive behaviors: The influence of halo effects. *Journal of Abnormal Child Psychology, 21*(5), 519–533. https://doi.org/10.1007/BF00916317

Bauer, J.-R., & Booth, A. E. (2019). Exploring potential cognitive foundations of scientific literacy in preschoolers: Causal reasoning and executive function. *Early Childhood Research Quarterly, 46*, 275–284. https://doi.org/10.1016/j.ecresq.2018.09.007

Cook, C., Goodman, N. D., & Schulz, L. E. (2011). Where science starts: Spontaneous experiments in preschoolers' exploratory play. *Cognition, 120*(3), 341–349. https://doi.org/10.1016/j.cognition.2011.03.003

Evans, N. S., Burke, R., Vitiello, V., Zumbrunn, S., & Jirout, J. J. (2023). Curiosity in classrooms: An examination of curiosity promotion and suppression in preschool math and science classrooms. *Thinking Skills and Creativity, 49*, 101333. https://doi.org/10.1016/j.tsc.2023.101333

Fetters, M. D., Curry, L. A., & Creswell, J. W. (2013). Achieving integration in mixed methods

designs—Principles and practices. *Health Services Research, 48*(6pt2), 2134–2156. https://doi.org/10.1111/1475-6773.12117

Gerde, H. K., Pierce, S. J., Lee, K., & Van Egeren, L. A. (2018). Early childhood educators' self-efficacy in science, math, and literacy instruction and science practice in the classroom. *Early Education and Development, 29*(1), 70–90. https://doi.org/10.1080/10409289.2017.1360127

Gerde, H. K., Schachter, R. E., & Wasik, B. A. (2013). Using the scientific method to guide learning: An integrated approach to early childhood curriculum. *Early Childhood Education Journal, 41*(5), 315–323. https://doi.org/10.1007/s10643-013-0579-4

Gopnik, A. (2012). Scientific thinking in young children: Theoretical advances, empirical research, and policy implications. *Science, 337*(6102), 1623–1627. https://doi.org/10.1126/science.1223416

Greenfield, D. B., Jirout, J., Dominguez, X., Greenberg, A., Maier, M., & Fuccillo, J. (2009). Science in the preschool classroom: A programmatic research agenda to improve science readiness. *Early Education and Development, 20*(2), 238–264. https://doi.org/10.1080/10409280802595441

Hartung, C. M., Van Pelt, J. C., Armendariz, M. L., & Knight, L. A. (2006). Biases in ratings of disruptive behavior in children: Effects of sex and negative halos. *Journal of Attention Disorders, 9*(4), 620–630. https://doi.org/10.1177/1087054705284233

Hsieh, H.-F., & Shannon, S. E. (2005). Three approaches to qualitative content analysis. *Qualitative Health Research, 15*(9), 1277–1288. https://doi.org/10.1177/1049732305276687

Jacobs, V. R., Lamb, L. L. C., & Philipp, R. A. (2010). Professional noticing of children's mathematical thinking. *Journal for Research in Mathematics Education, 41*(2), 169–202. https://doi.org/10.5951/jresematheduc.41.2.0169

Jirout, J., & Klahr, D. (2012). Children's scientific curiosity: In search of an operational definition of an elusive concept. *Developmental Review, 32*(2), 125–160. https://doi.org/10.1016/j.dr.2012.04.002

Kaiser, J., Retelsdorf, J., Südkamp, A., & Möller, J. (2013). Achievement and engagement: How student characteristics influence teacher judgments. *Learning and Instruction, 28*, 73–84. https://doi.org/10.1016/j.learninstruc.2013.06.001

Karacan, B., Aktar Songur, I., Ozaslan, A., & Iseri, E. (2026). *When rating scales fall short: LLM-assisted discovery of ADHD signals in Turkish teacher narratives* [Preprint]. arXiv. https://arxiv.org/abs/2606.02509

Koerber, S., Mayer, D., Osterhaus, C., Schwippert, K., & Sodian, B. (2015). The development of scientific thinking in elementary school: A comprehensive inventory. *Child Development, 86*(1), 327–336. https://doi.org/10.1111/cdev.12298

Larimore, R. A. (2020). Preschool science education: A vision for the future. *Early Childhood Education Journal, 48*(6), 703–714. https://doi.org/10.1007/s10643-020-01033-9

Legare, C. H. (2012). Exploring explanation: Explaining inconsistent evidence informs exploratory, hypothesis-testing behavior in young children. *Child Development, 83*(1), 173–185. https://doi.org/10.1111/j.1467-8624.2011.01691.x

Legare, C. H., Schult, C. A., Impola, M., & Souza, A. L. (2016). Young children revise explanations in response to new evidence. *Cognitive Development, 39*, 45–56.

https://doi.org/10.1016/j.cogdev.2016.03.003

Machts, N., Kaiser, J., Schmidt, F. T. C., & Möller, J. (2016). Accuracy of teachers' judgments of students' cognitive abilities: A meta-analysis. *Educational Research Review, 19*, 85–103. https://doi.org/10.1016/j.edurev.2016.06.003

Maier, M. F., Greenfield, D. B., & Bulotsky-Shearer, R. J. (2013). Development and validation of a preschool teachers' attitudes and beliefs toward science teaching questionnaire. *Early Childhood Research Quarterly, 28*(2), 366–378. https://doi.org/10.1016/j.ecresq.2012.09.003

Pendergast, E., Lieberman-Betz, R. G., & Vail, C. O. (2017). Attitudes and beliefs of prekindergarten teachers toward teaching science to young children. *Early Childhood Education Journal, 45*(1), 43–52. https://doi.org/10.1007/s10643-015-0761-y

Piekny, J., & Maehler, C. (2013). Scientific reasoning in early and middle childhood: The development of domain-general evidence evaluation, experimentation, and hypothesis generation skills. *British Journal of Developmental Psychology, 31*(2), 153–179. https://doi.org/10.1111/j.2044-835X.2012.02082.x

Saçkes, M., Trundle, K. C., Bell, R. L., & O'Connell, A. A. (2011). The influence of early science experience in kindergarten on children's immediate and later science achievement: Evidence from the Early Childhood Longitudinal Study. *Journal of Research in Science Teaching, 48*(2), 217–235. https://doi.org/10.1002/tea.20395

Schulz, L. E., & Bonawitz, E. B. (2007). Serious fun: Preschoolers engage in more exploratory play when evidence is confounded. *Developmental Psychology, 43*(4), 1045–1050. https://doi.org/10.1037/0012-1649.43.4.1045

Spektor-Levy, O., Kesner Baruch, Y., & Mevarech, Z. (2013). Science and scientific curiosity in pre-school—The teacher's point of view. *International Journal of Science Education, 35*(13), 2226–2253. https://doi.org/10.1080/09500693.2011.631608

Stiewe, A., & Ehm, J.-H. (2026). Preschool teachers' noticing: A review of conceptualizations, operationalization, and empirical relations. *Zeitschrift für Erziehungswissenschaft, 29*(2), 627–656. https://doi.org/10.1007/s11618-025-01343-x

Subhan, S., Puad, L. M. A. Z., & Listyowati, A. (2026). Assessing scientific thinking in early childhood: Cross-sectional evidence for a six-dimensional hierarchical structure in Indonesian preschoolers. *Al-Athfal: Jurnal Pendidikan Anak, 12*(1), 1–18. https://doi.org/10.14421/al-athfal.2026.121-01

Südkamp, A., Kaiser, J., & Möller, J. (2012). Accuracy of teachers' judgments of students' academic achievement: A meta-analysis. *Journal of Educational Psychology, 104*(3), 743–762. https://doi.org/10.1037/a0027627

Toplak, M. E., West, R. F., & Stanovich, K. E. (2013). Practitioner review: Do performance-based measures and ratings of executive function assess the same construct? *Journal of Child Psychology and Psychiatry, 54*(2), 131–143. https://doi.org/10.1111/jcpp.12001

van der Graaf, J., Segers, E., & Verhoeven, L. (2015). Scientific reasoning abilities in kindergarten: Dynamic assessment of the control of variables strategy. *Instructional Science, 43*(3), 381–400. https://doi.org/10.1007/s11251-015-9344-y

van Schijndel, T. J. P., Jansen, B. R. J., & Raijmakers, M. E. J. (2018). Do individual differences in children's curiosity relate to their inquiry-based learning? *International Journal of Science Education, 40*(9), 996–1015.

https://doi.org/10.1080/09500693.2018.1460772

Zimmerman, C. (2007). The development of scientific thinking skills in elementary and middle school. *Developmental Review, 27*(2), 172–223. https://doi.org/10.1016/j.dr.2006.12.001