

RESEARCH ARTICLE

Proposing a Generative AI-Based Jigsaw Instructional Model: A Case of Nature of Science Education for Science Teachers

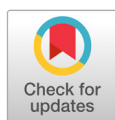
In-Gyu Kang¹, Seoung-Hey Paik^{2*}¹Sejong High School²Department of Science Gifted Education, Korea National University of Education

*Corresponding Author: Seoung-Hey Paik, shpaik@knue.ac.kr

ABSTRACT

Generative artificial intelligence (GenAI) often produces responses that are uncertain or factually incorrect. This raises the educational challenge of how learners can critically evaluate GenAI-generated responses rather than accept or reject them unreflectively. This study addressed this challenge through two cycles of action research in which science teachers used a GenAI-based Jigsaw (GJ) activity to explore the history of atomic theory (Dalton, Gay-Lussac, and Avogadro) for nature of science (NOS) education. In Cycle 1, five teachers completed the GJ activity. Their classroom discourse and reports revealed elaborated understanding of NOS. However, several design problems also emerged, including limited individual exploration time and uncritical acceptance of unfamiliar experts' explanations. These findings informed concrete refinements for Cycle 2, such as expanded individual exploration time, overlapping expert-area assignments, and explicit prompting guidance. The RFN framework was also adopted because it better captures the educational-application and domain-specific dimensions of GenAI-supported instruction. In Cycle 2, 25 chemistry teachers completed the refined GJ activity. Activity records showed teachers verifying GenAI responses through re-questioning, peer comparison, and researcher support, and pre-post RFN questionnaire results showed exploratory increases in three of six categories (Aims and Values, Scientific Knowledge, and Educational Applications). Based on the findings from both cycles and the Cycle 2 interaction analysis, we propose a five-step GJ learning model—Structuring, Individual Inquiry, Expert-Group Learning, Collaborative Verification, and Synthesis. In this model, GenAI functions as a tool for individual distributed cognition, and Jigsaw structures peer and instructor interaction to support socially distributed cognition.

Keywords: Generative artificial intelligence, Jigsaw, nature of science, distributed cognition, action research, science teacher education



OPEN ACCESS

Brain, Digital, & Learning
2026, Vol. 16, No. 3, 305–318

<https://doi.org/10.31216/BDL.2026.16.3.6>

Received: August 24, 2026

Revised: September 25, 2026

Accepted: September 28, 2026

© 2026 Institute of Brain·AI based Education,
Korea National University of Education



This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (<http://creativecommons.org/licenses/by-nc/4.0/>) which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited.

Introduction

Generative artificial intelligence (GenAI) is reshaping the generation, evaluation, and communication of scientific knowledge (Chen & Zou, 2025; Krenn et al., 2022), and its

educational use is expanding rapidly (Deranty & Corbin, 2024). However, even advanced models remain prone to hallucinations and factual errors. Recent advances in multi-agent cross-checking (Jamshidi, 2026), retrieval-augmented generation (Zhang & Zhang, 2025), and refined instruction-tuning (Si et al., 2025) have improved GenAI reliability, but these approaches reduce rather than eliminate such errors. Technical safeguards alone therefore remain insufficient, and learners still need to critically evaluate GenAI-generated responses through evidence-based interaction with others.

Distributed cognition offers a framework for designing such learning environments. Cognition is distributed across individuals, tools, peers, and the environment rather than confined to an individual mind (Hutchins, 2001). Individual distributed cognition involves tools that extend a person's cognitive capacity, whereas socially distributed cognition involves resources shared through social interaction (Kim & Bae, 2012). GenAI can extend individual cognition, but its susceptibility to bias and error limits independent use (Banh & Strobel, 2023). Learners need opportunities to articulate, compare, and evaluate GenAI responses with peers and instructors. GenAI-supported learning should therefore connect individual inquiry with socially distributed evaluation.

To apply this principle in instruction, learners need subject matter that allows them to interpret, justify, and communicate GenAI responses, as well as a structure for cooperative learning. NOS education provides an appropriate context for these processes. It addresses how scientific knowledge is generated and evaluated within individual scientific practice and how scientific communities reach social-institutional consensus on that knowledge through evidence, critical scrutiny, and interaction (Erduran & Dagher, 2014; Erduran et al., 2019; Lederman et al., 2002). The relationship between individual judgment and communal consensus in scientific practice parallels the movement from individual GenAI inquiry to socially distributed evaluation. For this reason, NOS is well suited to examining GenAI-supported learning from a distributed-cognition perspective.

Science teacher education is particularly important in examining this relationship. Teachers' understanding of NOS affects how they support students' NOS understanding (Abd-El-Khalick & Lederman, 2000). Teachers also need direct experience in evaluating GenAI responses through evidence and discussion before guiding students in similar activities.

Jigsaw cooperative learning can provide the required social structure. Jigsaw links expert-group inquiry with reciprocal teaching in a home group (Aronson, 2002). Learners must articulate interpretations, compare information, and evaluate claims with peers. Jigsaw has also shown a large effect (Cohen's $d = 1.91$) on learners' ability to evaluate the credibility of online information (Parsazadeh et al., 2018). GenAI, NOS, Jigsaw, and distributed cognition play complementary roles within the GJ activity. GenAI provides individualized and potentially divergent responses, while NOS serves as the conceptual focus through which these responses are examined and discussed. Jigsaw extends this individual examination into communication and comparison with peers, and distributed cognition provides a perspective for interpreting how these cognitive processes are distributed across learners, tools, peers, and instructors.

Previous studies have examined separate aspects of this integration. AI has been incorporated as a peer in Jigsaw cooperative learning (Lee et al., 2025). However, the AI speaker delivered preprogrammed responses to structured questions, and learners were not required to verify its content. GenAI has also been used for NOS learning (Richardson, 2025), but the chatbot activity was conducted individually and did not include cooperative communication or verification. GenAI-supported teaching and learning research has expanded across diverse science education contexts, including

lesson planning around classic experiments (Clark et al., 2024), instructional sequences and modeling activities in chemistry (Gonçalves Costa et al., 2024; Lee et al., 2026), biology education (Byeon & Kwon, 2023), and collaborative learning with preservice teachers (Jang, 2023). However, these studies did not provide a staged Jigsaw structure for collaboratively verifying variable GenAI responses. These limitations indicate the need for a staged design that links individual GenAI inquiry with expert-group discussion and collaborative, evidence-based verification in NOS teacher education.

The present study developed a GenAI-based Jigsaw (GJ) activity in which teachers examined, verified, and reconstructed GenAI responses through individual inquiry, expert-group discussion, and collaborative verification. The activity focused on the historical debate among Dalton, Gay-Lussac, and Avogadro over atomic and molecular theories. This topic required teachers to compare competing perspectives and justify claims rather than identify a predetermined correct answer. Through two cycles of action research, the study addressed two research questions. (a) How was the GJ activity structured and refined through its implementation with teachers? (b) How did teachers evaluate GenAI responses during the activity, and what possible changes in their NOS understanding accompanied the process? Based on the findings, the study proposes a distributed-cognition-based GJ learning model for cooperative GenAI-supported instruction.

Research Methods

Research Procedure

Action research is a process-oriented method in which researchers actively engage with participants (Lee et al., 2005). This approach enhances teachers' capacity by fostering professional development and increasing self-awareness (Johnson, 2012; Tran, 2009) and is typically implemented across two or three cycles (Kemmis et al., 2014), each involving planning, action, observation, and reflection (Lee et al., 2005). In education, it provides a reflective framework through which professionals evaluate and improve their practice (Hine, 2013). This study followed two such cycles with science teachers. In Cycle 1, five teachers completed the GJ activity as an exploratory, qualitative cycle. Their classroom discourse and reflective reports were analyzed to identify strengths and limitations of the initial design, which informed the refinements carried into Cycle 2 (Table 2). In Cycle 2, the refined GJ activity was implemented with 25 science teachers using a mixed-methods design to examine possible changes in teachers' understanding of NOS.

Research Participants

This study was approved by the Institutional Review Board of Korea National University of Education (Approval No. KNUE-2023-09-SB-0337-01). All participants were informed of the research purpose and procedures and voluntarily agreed to participate. For Cycle 1, five in-service teachers enrolled in a graduate program completed the GJ activity over six 4-hour sessions using ChatGPT-3.5. Only one had previously used GenAI in the classroom. For Cycle 2, in-service chemistry teachers were recruited from the same graduate program, excluding Cycle 1 participants. Of these,

25 (8 middle school, 17 high school) completed the cycle, which was conducted over four 4-hour sessions using GPT-4o mini. Most Cycle 2 teachers (20 of 25, 80%) had used ChatGPT before, but most used it infrequently (15 of 25 less than once a month) due to concerns about response errors and hallucination. Cycle 2 teachers' teaching experience was distributed across three bands: 15 years or more (8 of 25, 32%), 11–14 years (7 of 25, 28%), and 10 years or fewer (10 of 25, 40%). In both cycles, teachers received guidance on prompting and practiced using GenAI before the GJ activity began.

GJ Activity

The GJ activity extended the Jigsaw cooperative structure (Aronson, 2002) by embedding GenAI as a tool for building expert knowledge and integrating a three-stage verification structure—individual, expert-group, and collaborative—into Jigsaw's four steps Table 1. Prompting and response verification were treated as related but distinct aspects of GenAI use. Prompting involved formulating and refining queries, whereas verification involved evaluating consistency, evidential support, and historical plausibility of responses. Thus, the preliminary prompting guidance in Cycle 2 provided procedural support for GenAI use rather than establishing verification competence. In Step 1, the researcher assigned teachers to home groups and, from Cycle 2, to two overlapping expert perspectives among Dalton, Gay-Lussac, and Avogadro—addressing Cycle 1's single-expert assignment. Teachers then individually used GenAI to explore their assigned perspectives and evaluated its responses. In Step 2, teachers sharing an expert perspective compared their individually gathered GenAI responses, discussed their consistency, and jointly constructed expert-area materials. In Step 3, teachers returned to their home groups and shared these expert-area materials. Unlike Steps 1 and 2, this step did not involve GenAI, and home-group members collaboratively reviewed and cross-verified the shared materials among themselves. In Step 4, teachers individually completed a piece of scientific writing integrating their understanding, on which the researcher provided feedback grounded in NOS. Based on the findings from both cycles, the Results and Discussion section below reorganizes this three-stage verification structure into a five-step model.

Table 1. The GJ activity in Cycle 2 (adapted from Kang, 2026, Table III-7).

Step	Description	GenAI Used
1. Home-Group & Expert Assignment	Teachers are assigned to overlapping expert perspectives (two of three—Dalton, Gay-Lussac, Avogadro) and individually use GenAI to explore and evaluate responses in their expert area	Yes
2. Expert-Group Learning	Teachers sharing an expert perspective compare their GenAI responses and jointly construct expert-area materials	Yes
3. Home-Group Sharing & Review	Teachers share expert-area materials with their home group. Members collaboratively review and cross-verify the materials	No
4. Individual Evaluation	Teachers complete individual scientific writing integrating their understanding, with researcher feedback	No

Data Collection and Analysis

In Cycle 1, GJ sessions were conducted online via Zoom, video-recorded, and transcribed using Clova Note. Teachers' GenAI interaction logs were collected through a shared-chat function, and post-activity reflective reports were collected via Google Forms. Teachers' understanding of NOS was analyzed using the Views of Nature of Science (VNOS) framework (Abd-El-Khalick & Lederman, 2000; Lederman et al., 2002), which examines conceptions of the empirical

nature of scientific knowledge, the relationship between theories and laws, the tentativeness of scientific knowledge, creativity and imagination, and the sociocultural embeddedness of science.

The unit of analysis was each teacher utterance in the session discourse and each statement in the post-activity reports that contained a NOS-relevant claim. These statements were categorized according to the five VNOS categories. The two researchers first examined the data independently and then compared their coding. Disagreements were discussed until consensus was reached. To enhance credibility, the researchers triangulated the interpretations across session recordings, GenAI interaction logs, and post-activity reports. A statement was interpreted as demonstrating an elaborated understanding when the teacher supported a NOS-related claim with an explicit justification, comparison, or explanation grounded in the historical case.

In Cycle 2, teachers completed the Reconceptualized Family Resemblance Approach to NOS Questionnaire (RFNQ) before and after the activity. The RFNQ organizes NOS into six categories across three dimensions—a cognitive-epistemological dimension (Aims and Values, Scientific Practices, Scientific Methods, Scientific Knowledge), a social-institutional dimension (Social-Institutional Aspects), and an educational-application dimension (Educational Applications) (Kaya et al., 2019). The English RFNQ was translated into Korean, and both versions were provided to participants to prevent misunderstanding. The Korean translation was produced through consensus among NOS science education experts and independently verified by the researchers. As in Cycle 1, sessions were recorded via Zoom and transcribed with Clova Note, GenAI interaction logs from the Expert-Group step were collected through a shared function, expert-group materials were collected via Canva, and whole-group presentation materials and post-activity reports were collected online. No individual interviews were conducted in Cycle 2.

During GJ activity, the researcher provided prompting guidance, used follow-up questions to support teachers when they encountered conceptual difficulties in interpreting GenAI responses, and provided relevant historical context when further clarification or verification was needed. Accordingly, researcher–teacher interactions were coded separately and treated as part of the instructional context.

Qualitative data therefore comprised classroom and discussion transcripts, GenAI interaction records, and written reports rather than interview data. For quantitative analysis, RFNQ responses were scored on a 5-point Likert scale, with reverse scoring applied to the relevant items. Given the multidimensional structure of the RFNQ, internal consistency across the three RFN dimensions was assessed primarily using McDonald's ω , with Cronbach's α as a supplementary measure. Normality was assessed using the Shapiro–Wilk test. Paired-samples *t* tests or Wilcoxon signed-rank tests were conducted as appropriate. Effect sizes were calculated for each RFN category using rank-biserial *r* for the Wilcoxon tests and Cohen's *d* as a supplementary indicator. The six RFN category-level analyses were treated as exploratory because the categories are conceptually interconnected (Cheung & Erduran, 2023). Accordingly, *p* values were not adjusted for multiple comparisons.

For qualitative analysis, activity records and reports were analyzed according to RFN category. Interactions among teachers, GenAI, peers, and the researcher were coded according to interaction pair (e.g., teacher–GenAI and teacher–group) based on transcripts of verbal exchanges generated using Clova Note. Verification referred to episodes in which teachers examined an existing GenAI-generated claim by re-questioning GenAI, comparing responses with peers, consulting external sources, or seeking researcher clarification, rather than simply extending the conversation.

Verification behavior itself was not treated as evidence of NOS understanding. A verification episode was considered relevant to NOS only when the discussion addressed issues represented in the VNOS or RFN frameworks. These included evidence, scientific methods, models and theories, and the social evaluation and acceptance of scientific knowledge.

The coded interactions were interpreted through the lens of distributed cognition to characterize how GenAI responses were compared and evaluated during individual inquiry and social interaction. The two researchers independently coded the data, identified illustrative episodes in which teachers evaluated GenAI responses, and classified these episodes by RFN category. Coding discrepancies were resolved through discussion. Triangulation across data sources was used to enhance the credibility of the interpretations.

Results and Discussion

Cycle 1: Findings and Refinements for Cycle 2

Prior to the Jigsaw structure, teachers' individual GenAI exploration on a shared topic (Raoult's law) surfaced frequent response errors and uneven depth and direction of inquiry, indicating that individual use alone did not support the shared justification of knowledge NOS instruction requires. The researcher therefore guided teachers through the GJ activity and implemented it.

Within the GJ activity, teachers evaluated GenAI responses in several ways—re-questioning GenAI or drawing on their own expertise, comparing responses against peers' findings or internet sources, or relying on the researcher's explanations. For example, when the bead metaphor in a GenAI response for molecular volume confused teachers about particle spacing, the researcher's clarification resolved the misunderstanding. This comparison and verification gave teachers occasion to critically review and negotiate scientific concepts.

Cycle 1 also surfaced limitations. In Step 2, GenAI's broad, diverse responses strained discussion time. In Step 3, lacking background knowledge outside their own expert area, teachers sometimes accepted other experts' explanations with limited critical questioning. Teachers described needing the researcher to supply background knowledge and activate communication about GenAI responses rather than only observe, positioning the instructor as integral to the activity's verification process.

Analysis of the five teachers' Cycle 1 discourse and reports using the VNOS framework showed elaborated understanding across all five dimensions Table 2. For example, examining Dalton's atomic theory against the technological limits of his era, one teacher moved from describing scientific knowledge as fixed fact toward describing it as constructed through scientists' experience and reasoning. Comparing Avogadro's and Dalton's differing conclusions from related evidence, other teachers described scientific models as provisional tools rather than absolute representations. Learning that Dalton, Gay-Lussac, and Avogadro were contemporaries, teachers described the way their achievements were recognized as reflecting a need for social consensus rather than the order in which knowledge emerged. The excerpts in Table 2 were selected as representative examples of how each VNOS category was expressed in the Cycle 1 data and are not intended to represent every participant. Participants did not articulate NOS-related understanding in

the same way, and some teachers experienced conceptual confusion or required researcher support during NOS-related discussions.

Table 2. Teachers' VNOS understanding in Cycle 1

VNOS Category	Understanding Demonstrated	Illustrative Excerpt
Empirical nature of scientific knowledge	Science described as constructed through scientists' experience and reasoning, not only fixed facts	Science becomes more solid and alive through human experience and thinking
Theories and laws	Theories and laws distinguished as different but complementary types of knowledge	Distinguished Gay-Lussac's law from Dalton's and Avogadro's theories
Tentativeness of scientific knowledge	Models described as provisional tools for explaining phenomena, not absolute truths	A model is created to explain certain phenomena
Creativity and imagination	Creativity described as essential across scientists' differing theoretical approaches	Explained Avogadro's use of mathematical reasoning to resolve Dalton's atomic theory
Sociocultural embeddedness	Scientific acceptance described as requiring social and institutional consensus	Avogadro's theory needed more effort toward consensus among scientists

These findings and observations informed the Cycle 2 design refinements in Table 3. In Step 2, teachers reported that extensive GenAI exploration left insufficient time for verification and discussion. In Step 3, limited knowledge of other expert areas sometimes led teachers to accept peers' explanations with little scrutiny. As one teacher noted, "Since we didn't know much about the other areas, we tended to accept the material rather than question it." Teachers also requested guidance on prompting strategies. These observations informed expanded exploration time, overlapping expert-area assignments, and prompting-strategy guidance. The researchers also designated specific platforms for activity outputs and adopted RFN in place of VNOS.

Table 3. Design refinements from Cycle 1 to Cycle 2

Issue Identified in Cycle 1	Refinement for Cycle 2
Limited time for individual GenAI exploration in Step 2	Expansion of GenAI-supported individual inquiry in Step 1
Teachers readily accepted unfamiliar experts' explanations	Overlapping expert-area assignments (each teacher assigned two of the three scientists)
Activity outputs were scattered across unconsolidated platforms (Notion, Padlet, Canva)	GenAI response sharing via Google Sheets and output consolidation in Canva
Teachers varied in prompting skill	Advance presentation of key prompting strategies with examples
VNOS did not capture the educational-application and domain-specific aspects of GenAI-supported instruction	Adoption of the RFN framework for domain-specific analysis and Educational Applications

Cycle 2: The GJ Activity and Teachers' NOS Understanding

In Cycle 2, the refined GJ activity ran across four 4-hour sessions—an initial session on prompting strategies (chain-of-thought and few-shot techniques), one session each for Steps 1 and 2, and a final session combining Steps 3 and 4. Twenty-five teachers formed six expert groups, each assigned two of the three scientists. The interaction pairs differed in their main topics. Teacher–GenAI exchanges were most frequent (70, 29.5%) and focused on exploratory questioning and verification. Group–group exchanges (55, 23.2%) focused on expert-knowledge transfer and collaborative verification, while researcher–teacher exchanges (33, 13.9%) focused on verification and communication support. Researcher–GenAI exchanges (29, 12.2%) focused on pre-verification and guidance on GenAI use. Teacher–group

and teacher–resource exchanges (25 each, 10.5%) focused on exploratory dialogue and fact-checking, respectively. Overall, group and researcher interactions were mainly associated with verification and knowledge transfer, whereas teacher–GenAI exchanges were more closely associated with exploratory inquiry. Analysis of the activity records identified recurring teacher–GenAI, researcher–teacher, teacher–group, and group–group interaction patterns. Before teachers began questioning GenAI in Step 1, the researcher offered prompting guidance, such as suggesting that teachers probe Gay-Lussac's work through the more fully documented Dalton and Avogadro. In Step 1, teachers individually questioned GenAI 5–23 times ($M = 13.96$) while exploring their assigned perspectives and re-questioned responses when inconsistencies emerged. In Step 2, teachers compared GenAI-generated claims with peers' responses and external evidence, including a chemistry textbook, and revised their interpretations when discrepancies were identified. The researcher also provided historical context when needed, for example by noting that Dalton, Gay-Lussac, and Avogadro were contemporaries to prompt further questioning. In Step 3, home groups cross-checked the accounts brought from different expert groups and reconsidered or revised interpretations when inconsistencies remained. Within the GJ activity, teachers' evaluation, revision, and integration of GenAI responses drew on GenAI, peers, researcher support, and external evidence, indicating that cognitive resources were distributed across multiple sources.

Internal consistency, evaluated across the three RFN dimensions, was acceptable in both pre- and post-tests (McDonald's $\omega = .77$ – $.88$). Shapiro-Wilk tests indicated that Scientific Practices and Social-Institutional Aspects met the assumption of normality ($p = .069$ and $p = .997$, respectively), while the remaining four categories did not. Paired t-tests were therefore applied to the former two categories and Wilcoxon signed-rank tests to the latter four. Scientific Practices and Social-Institutional Aspects showed no statistically significant pre-post difference ($t(24) = 0.99$, $p = .331$; $t(24) = 1.40$, $p = .174$). Among the categories analyzed using Wilcoxon signed-rank tests, Aims and Values ($z = -2.57$, $p = .010$, $r = .81$), Scientific Knowledge ($z = -2.05$, $p = .040$, $r = .53$), and Educational Applications ($z = -2.12$, $p = .034$, $r = .47$) showed exploratory pre-post increases with unadjusted p values below .05. The percentages of teachers whose scores increased were 36%, 44%, and 52%, respectively. Scientific Methods showed no significant change ($z = -1.42$, $p = .156$). As a supplementary indicator, paired-samples Cohen's d for all six categories ranged from 0.16 to 0.41, corresponding to small-to-medium effects. Pre-post RFN questionnaire scores are summarized in Table 4.

Table 4. Pre-post RFN scores by category in Cycle 2 (adapted from Kang, 2026, Table IV-7)

RFN Category	Pre M (SD)	Post M (SD)	Test Statistic	p	Effect Size(d)
Aims and Values	26.60 (2.57)	27.12 (2.35)	$z = -2.57$	.010	0.41
Scientific Practices	50.96 (4.68)	51.40 (4.50)	$t(24) = 0.99$	.331	0.16
Scientific Methods	27.44 (3.70)	27.80 (3.39)	$z = -1.42$	.156	0.25
Scientific Knowledge	32.44 (3.06)	32.96 (2.84)	$z = -2.05$	.040	0.33
Social-Institutional Aspects	63.52 (7.02)	64.80 (5.53)	$t(24) = 1.40$	.174	0.23
Educational Applications	64.24 (6.29)	65.92 (5.41)	$z = -2.12$	.034	0.33

Teachers participating in this study had generally higher baseline RFNQ scores than UAE preservice and early-years teachers but showed the same relative weakness in Scientific Methods (Schofield et al., 2023; Takriti et al., 2023). The exploratory analysis showed a pre-post increase in Educational Applications, with an unadjusted p value below .05.

Qualitative analysis of Step 4 reports illustrates the NOS understanding underlying these RFNQ gains, organized

by RFN category. In Aims and Values, teachers linked scientists' philosophical dispositions to their theoretical choices. For example, one teacher wrote that Gay-Lussac “had strong experimental skill and verified data accurately, but was philosophically conservative, which limited his ability to broaden theory” (Step 4 report). In Scientific Knowledge, teachers described models as provisional and complementary rather than competing. For instance, one teacher noted during a Cycle 2 expert-group discussion that Avogadro “created the molecule concept to mathematically explain Gay-Lussac's law, further extending the meaning of atomic theory,” and another wrote that “a model is only a tool for explaining scientific knowledge, not something perfect” (Step 4 report). In Educational Applications, teachers connected the activity to their own teaching. In another case, one teacher, who had assumed atoms, elements, and molecules were “too familiar to hold anything new,” reported that exploring the scientists' historical context “made me want to understand it more deeply” (Step 4 report), and another proposed re-enacting the Einstein–Bohr debate with ChatGPT so students could learn that scientific knowledge “develops provisionally through debate and criticism, rather than as absolute truth” (Step 4 report).

Verification became relevant to NOS when teachers moved beyond simply judging whether a GenAI response was correct and discussed NOS-related issues embedded in the response. For example, teachers questioned GenAI's characterization of Avogadro's reasoning as deductive and discussed whether it was better understood as abductive, thereby examining how scientific methods are characterized.

Beyond these RFNQ-linked gains, analysis of the coded teacher–GenAI interactions illustrates how teachers evaluated GenAI responses across the activity's three verification stages. In Step 1, teachers re-questioned GenAI when its responses were internally inconsistent. For example, after GenAI gave conflicting answers about whether Avogadro's theory was well received by his contemporaries, one teacher revised the prompt to resolve the contradiction. In Step 2, expert groups checked GenAI's historical claims against each other. For instance, when GenAI stated that Gay-Lussac had sent Dalton a letter, the group questioned its factual basis until the researcher confirmed the exchange had occurred. In another group, a teacher challenged a peer's GenAI-sourced claim that Dalton had supported Gay-Lussac's objections by citing a separate GenAI response describing Dalton's own explanation of combination reactions. A further case involved external evidence. A teacher shared pages from a chemistry textbook showing that Dalton had explained the reaction through caloric theory rather than as described by GenAI, and the group revised its initial interpretation accordingly. Because this claim was evaluated against evidence outside the GenAI interaction, the episode represents verification beyond comparison among variable GenAI responses. In Step 3, home groups cross-checked GenAI responses gathered by different expert groups. In one case, when GenAI attributed an inductive approach to Gay-Lussac and a deductive approach to Avogadro, teachers agreed the two scientists used different methods but judged GenAI's specific attribution as inaccurate, noting that Gay-Lussac's experimentally grounded law and Avogadro's theoretical construction did not map cleanly onto that distinction. Across these stages, teachers verified GenAI responses through their own historical knowledge, peer comparison, and researcher clarification.

A Distributed-Cognition-Based GJ Model

The findings from the two action-research cycles led to the proposal of a GJ learning model comprising five functionally distinct steps: Structuring, Individual Inquiry, Expert-Group Learning, Collaborative Verification, and

Synthesis. The proposed model is presented in Fig. 1.

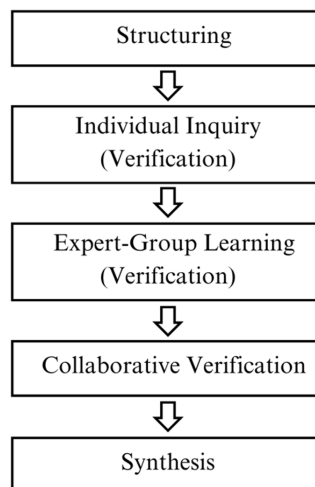


Fig. 1. A distributed-cognition-based GJ learning model (adapted from Kang, 2026, Figure IV-3)

This five-step formulation does not introduce procedures beyond those implemented in the study, nor does it represent an empirically validated sequence. Rather, it is proposed as a design conjecture that makes explicit the different instructional and cognitive functions that became visible through the two cycles.

Structuring was distinguished from Individual Inquiry because the allocation of expert perspectives served a different function from teachers' subsequent use of GenAI. In Cycle 1, teachers sometimes accepted explanations from unfamiliar expert areas with limited critical questioning. Cycle 2 therefore introduced overlapping expert assignments so that teachers had partial knowledge of more than one perspective before returning to the home group. Structuring thus concerned the distribution of perspectives and cognitive resources across teachers, whereas Individual Inquiry involved teachers' direct interaction with GenAI to generate, examine, and re-question information. The predominance of teacher–GenAI exchanges around exploratory questioning and verification is consistent with this interpretation of Individual Inquiry as primarily tool-mediated cognition.

Expert-Group Learning extended this individual inquiry into socially distributed comparison and verification. Teachers brought individually generated GenAI responses into expert groups, compared them with peers' responses, and drew on researcher support and external resources when evaluating discrepancies. The interaction patterns identified in Cycle 2 support this interpretation: group-level exchanges centered on expert-knowledge transfer and collaborative verification, while researcher–teacher interactions primarily supported verification and communication. External resources also contributed to fact-checking when teachers compared GenAI-generated claims with documentary evidence. Thus, expert-group activity did not simply transmit information generated during individual inquiry; it provided a setting in which those representations could be compared, challenged, and revised.

Collaborative Verification involved a further shift from within-perspective comparison to cross-perspective evaluation in the home group. Teachers communicated the accounts developed in different expert groups and reconsidered claims when inconsistencies remained across perspectives. This distinction is important because Cycle 1 showed that home-

group sharing could become passive acceptance when members lacked sufficient knowledge of other expert areas. The overlapping assignments introduced in Cycle 2 were intended to reduce this limitation by creating partial redundancy across expert perspectives and increasing opportunities for cross-checking. In this sense, the Jigsaw structure expanded GenAI-supported inquiry from individual teacher–tool interaction to socially distributed evaluation across teachers, peers, the researcher, and external resources.

Synthesis was distinguished from Collaborative Verification by the shift from group-level learning to individual reconstruction. In the final step, each teacher produced a written account integrating the perspectives examined in the preceding steps, and researcher feedback was directed to this individual product. Synthesis therefore represents the reorganization of information that had been distributed across GenAI, peers, and external resources into an individual account. The overall sequence can accordingly be interpreted as a movement from the distribution of perspectives, through individual and socially distributed inquiry and verification, to individual synthesis.

These distinctions also clarify how the proposed GJ model differs from traditional Jigsaw. In traditional Jigsaw, learners in the same expert area generally work from common instructional materials, whereas GJ learners construct expert knowledge through their own GenAI interactions and may therefore bring different responses even to the same expert perspective. GJ also makes evaluation of these variable responses explicit across individual, expert-group, and home-group activity. GenAI and cooperative interaction therefore operate jointly: GenAI expands the range of representations available for inquiry, while peers, the researcher, and external evidence provide additional cognitive resources for evaluating and reconstructing them. Accordingly, the five-step model is proposed from a distributed-cognition perspective as an exploratory design conjecture derived from the two action-research cycles.

Conclusions and Implications

This study developed a GenAI-based Jigsaw (GJ) activity for science teachers across two action-research cycles and examined possible changes in teachers' NOS understanding. Structuring GenAI use within a cooperative Jigsaw sequence, rather than unstructured individual use, was associated with elaborated understanding of the empirical, tentative, and sociocultural dimensions of science in Cycle 1. Cycle 2 showed exploratory pre-post increases in Aims and Values, Scientific Knowledge, and Educational Applications.

The findings have implications for three aspects of science education. When addressed through cooperative learning, GenAI hallucinations and errors can provide material for critical evaluation rather than being treated only as technical problems. The proposed model also presents a design conjecture for applying distributed cognition to instruction by combining GenAI as an individual cognitive tool with peer and instructor interaction for socially distributed evaluation. Teachers' direct participation in this form of inquiry may inform professional development for the critical use of GenAI before similar activities are introduced to students.

This study has several limitations. Because it employed a single-group pre-post design, maturation, testing effects, and the researcher's influence cannot be ruled out. The observed changes therefore cannot be attributed solely to the GJ activity. The Cycle 2 sample was also small ($n = 25$) for six uncorrected inferential comparisons, and the results should be interpreted as exploratory. In addition, the activity addressed only one content domain, the history of atomic theory,

leaving its transferability to other domains untested. Future research should examine the GJ model in other domains and include a comparison group.

The two cycles used ChatGPT-3.5 and GPT-4o mini, and differences in model version and probabilistic generation limit the exact reproducibility of the verification patterns observed. However, the GJ model is intended to use GenAI as a tool for individual distributed cognition and extend this cognition through Jigsaw-based social interaction, rather than depend on a particular model's error profile. Future studies should examine this rationale with newer GenAI models and repeated implementations.

From a distributed-cognition perspective, further studies should also investigate the cognitive roles of GenAI and other cognitive resources across processes of information generation, verification, revision, and integration.

Acknowledgements

This article is based in part on the first author's doctoral dissertation. We are grateful to the graduate students and teachers who participated in the study and provided useful feedback and research materials.

Conflicts of Interest

The authors declared no conflicts of interest.

References

- Abd-El-Khalick, F., & Lederman, N. G. (2000). The influence of history of science courses on students' views of nature of science. *Journal of Research in Science Teaching*, 37(10), 1057–1095. [https://doi.org/10.1002/1098-2736\(200012\)37:10<1057::AID-TEA3>3.0.CO;2-C](https://doi.org/10.1002/1098-2736(200012)37:10<1057::AID-TEA3>3.0.CO;2-C)
- Aronson, E. (2002). Building empathy, compassion, and achievement in the jigsaw classroom. In J. Aronson (Ed.), *Improving academic achievement: Impact of psychological factors on education* (pp. 209–225). Academic Press. <https://doi.org/10.1016/B978-012064455-1/50013-0>
- Banh, L., & Strobel, G. (2023). Generative artificial intelligence. *Electronic Markets*, 33(1), 63. <https://doi.org/10.1007/s12525-023-00680-1>
- Byeon, J. H., & Kwon, Y. J. (2023). An investigation of generative AI in educational application: Focusing on the usage of ChatGPT for learning biology. *Brain, Digital, & Learning*, 13(1), 1–17. <https://doi.org/10.31216/BDL.20230001>
- Chen, Y., & Zou, J. (2025). Simple and effective embedding model for single-cell biology built from ChatGPT. *Nature Biomedical Engineering*, 9(4), 483–493. <https://doi.org/10.1038/s41551-024-01284-6>
- Cheung, K. K. C., & Erduran, S. (2023). A systematic review of research on family resemblance approach to nature of science in science education. *Science & Education*, 32(5), 1637–1673. <https://doi.org/10.1007/s11191-022-00379-3>
- Clark, T. M., Phaner, M., Stoltzfus, M., & Queen, M. S. (2024). Using ChatGPT to support lesson planning for the historical experiments of Thomson, Millikan, and Rutherford. *Journal of Chemical Education*, 101(5), 1992–1999. <https://doi.org/10.1021/acs.jchemed.4c00200>
- Deranty, J. P., & Corbin, T. (2024). Artificial intelligence and work: A critical review of recent research from

- the social sciences. *AI & Society*, 39(2), 675–691. <https://doi.org/10.1007/s00146-022-01496-x>
- Erduran, S., & Dagher, Z. R. (2014). *Reconceptualizing the nature of science for science education: Scientific knowledge, practices and other family categories*. Springer. <https://doi.org/10.1007/978-94-017-9057-4>
- Erduran, S., Dagher, Z. R., & McDonald, C. V. (2019). Contributions of the family resemblance approach to nature of science in science education: A review of emergent research and development. *Science & Education*, 28, 311–328. <https://doi.org/10.1007/s11191-019-00052-2>
- Gonçalves Costa, G., Nascimento Júnior, W. J. D., Mombelli, M. N., & Giroto Júnior, G. (2024). Revisiting a teaching sequence on the topic of electrolysis: A comparative study with the use of artificial intelligence. *Journal of Chemical Education*, 101(8), 3255–3263. <https://doi.org/10.1021/acs.jchemed.4c00456>
- Hine, G. S. (2013). The importance of action research in teacher education programs. *Issues in Educational Research*, 23(2), 151–163. <https://search.informit.org/doi/10.3316/informit.354483461821862>
- Hutchins, E. (2001). Cognition, distributed. In N. J. Smelser & P. B. Baltes (Eds.), *International encyclopedia of the social & behavioral sciences* (pp. 2068–2072). Elsevier. <https://doi.org/10.1016/B0-08-043076-7/01636-3>
- Jamshidi, S. (2026). Collective hallucination in multi-agent LLMs: Modeling and defense. *arXiv*. <https://doi.org/10.48550/arXiv.2606.07941>
- Jang, E. J. (2023). Evaluation of affective domains and technology acceptability of collaborative learning using generative AI ChatGPT of preservice teachers. *Brain, Digital, & Learning*, 13(3), 229–243. <https://doi.org/10.31216/bdl.20230014>
- Johnson, A. P. (2012). *A short guide to action research* (4th ed.). Allyn and Bacon.
- Kang, I. G. (2026). *Proposing a generative AI-based Jigsaw learning model for nature of science education* [Doctoral dissertation, Korea National University of Education]. KNUE dCollection. <https://dcollection.knue.ac.kr/common/orgView/000000043642>
- Kaya, E., Erduran, S., Aksoz, B., & Akgun, S. (2019). Reconceptualised family resemblance approach to nature of science in pre-service science teacher education. *International Journal of Science Education*, 41(1), 21–47. <https://doi.org/10.1080/09500693.2018.1529447>
- Kemmis, S., McTaggart, R., & Nixon, R. (2014). *The action research planner: Doing critical participatory action research*. Springer. <https://doi.org/10.1007/978-981-4560-67-2>
- Kim, S. K., & Bae, J. H. (2012). A learning flow model supporting distributed cognition in IT education. *Journal of Convergence Security*, 12(6), 51–59. <https://scienceon.kisti.re.kr/srch/selectPORSrchArticle.do?cn=JAKO201206737754383>
- Krenn, M., Pollice, R., Guo, S. Y., Aldeghi, M., Cervera-Lierta, A., Friederich, P., dos Passos Gomes, G., Häse, F., Jinich, A., Nigam, A., Yao, Z., & Aspuru-Guzik, A. (2022). On scientific understanding with artificial intelligence. *Nature Reviews Physics*, 4(12), 761–769. <https://doi.org/10.1038/s42254-022-00518-3>
- Lederman, N. G., Abd-El-Khalick, F., Bell, R. L., & Schwartz, R. S. (2002). Views of nature of science questionnaire: Toward valid and meaningful assessment of learners' conceptions of nature of science. *Journal of Research in Science Teaching*, 39(6), 497–521. <https://doi.org/10.1002/tea.10034>
- Lee, G. G., Mun, S., Shin, M. K., & Zhai, X. (2025). Collaborative learning with artificial intelligence speakers: Pre-service elementary science teachers' responses to the prototype. *Science & Education*, 34(2), 847–875. <https://doi.org/10.1007/s11191-024-00526-y>
- Lee, H. G., Kim, S., & Paik, S. H. (2026). Exploring learning changes through generative AI-based simulation construction activities for acid–base equilibrium: A focus on academic achievement and STEAM competencies. *Brain, Digital, & Learning*, 16(2), 147–163. <https://doi.org/10.31216/bdl.2026.16.2.3>
- Lee, Y. S., Kim, Y. C., Lee, H. G., Kim, Y. M., Jo, D. J., & Jo, J. S. (2005). *Action research methods*. Hakjisa.

- Parsazadeh, N., Ali, R., & Rezaei, M. (2018). A framework for cooperative and interactive mobile learning to improve online information evaluation skills. *Computers & Education*, 120, 75–89. <https://doi.org/10.1016/j.compedu.2018.01.010>
- Richardson, C. (2025). The effects of AI on student understanding of NOS concepts: Science is tentative, is creative, and is affected by society and culture. In I. S. Carter, V. L. Akerson, & G. A. Buck (Eds.), *Exploring social and cultural contexts of the nature of science through action research* (pp. 191–219). ISTES Organization. <https://files.eric.ed.gov/fulltext/ED678652.pdf>
- Schofield, L., Takriti, R., Rabbani, L., AlAmirah, I., Ioannidou, O., Alhosani, N., Tairab, H., & Erduran, S. (2023). Early years education teachers' perceptions of nature of science. *International Journal of Science Education*, 45(8), 613–635. <https://doi.org/10.1080/09500693.2023.2168139>
- Si, S., Zhao, H., Chen, G., Gao, C., Bai, Y., Wang, Z., An, K., Luo, K., Qian, C., Qi, F., Chang, B., & Sun, M. (2025). Aligning large language models to follow instructions and hallucinate less via effective data filtering. *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 16469–16488. <https://aclanthology.org/2025.acl-long.804/>
- Takriti, R., Tairab, H., Alhosani, N., Elhoweris, H., Schofield, L., Rabbani, L., & AlAmirah, I. (2023). Toward understanding science as a whole: Investigating preservice teachers' perceptions about nature of science in the United Arab Emirates. *Science & Education*, 32(5), 1321–1361. <https://doi.org/10.1007/s11191-022-00404-5>
- Tran, T. T. H. (2009). Why is action research suitable for education? *VNU Journal of Foreign Studies*, 25(2), 97–106. <https://js.vnu.edu.vn/FS/article/view/2240>
- Zhang, W., & Zhang, J. (2025). Hallucination mitigation for retrieval-augmented large language models: A review. *Mathematics*, 13(5), 856. <https://doi.org/10.3390/math13050856>

Authors' Information

- Kang, In-Gyu: Sejong High School, Teacher, First-Author. <https://orcid.org/0009-0006-1362-7031>
- Paik, Seoung-Hey: Department of Science Gifted Education, Korea National University of Education, Professor, Corresponding Author. <https://orcid.org/0000-0002-3822-3838>