RESEARCH ARTICLE

Exploring Learning Changes through Generative AI–Based Simulation Construction Activities for Acid–Base Equilibrium: A Focus on Academic Achievement and STEAM Competencies

Hwang-Gi Lee[1], Sungki Kim[2*], Seoung-Hey Paik[3*]
[1]Gyeongsan Science High School
[2]Korea Institute for Curriculum and Evaluation
[3]Korea National University of Education

[*]Corresponding Author: Sungki Kim, mcarey2000@kice.re.kr; Seoung-Hey Paik, shpaik@knue.ac.kr

ABSTRACT

This study explored the potential of generative artificial intelligence–based simulation construction activities to facilitate changes in students' academic achievement and STEAM competencies in chemistry. The study was conducted with 54 second-year students enrolled in an advanced chemistry course at a science high school in Gyeongsangbuk-do, Republic of Korea. With the support of ChatGPT, students constructed particle-level simulations based on VPython and engaged in inquiry activities involving iterative modification and refinement of their simulations. Academic achievement was measured using five items related to acid–base equilibrium administered before and after the intervention, and STEAM competencies were assessed using a 21-item, four-point Likert-scale questionnaire. The results indicated that academic achievement scores increased from pretest to posttest. Differences across achievement levels were also observed, with students in the higher-achievement group showing greater gains. STEAM competencies also showed an overall increase in mean scores, and the largest change was observed in creative problem-solving. These findings suggest that the use of generative artificial intelligence in chemistry instruction may be related to context-specific changes in students' conceptual understanding and competencies, and provide implications for future research in generative artificial intelligence–supported chemistry education.

Keywords: Generative artificial intelligence, acid–base equilibrium, simulation construction, academic achievement, STEAM competencies

Check for updates

OPEN ACCESS

Brain, Digital, & Learning
2026, Vol. 16, No. 2, 147–163
https://doi.org/10.31216/BDL.2026.16.2.3

Received: April 20, 2026
Revised: June 30, 2026
Accepted: June 30, 2026

© 2026 Institute of Brain 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

With the rapid advancement of science and technology, digital technologies—including artificial intelligence (AI)—are increasingly demanding a restructuring of instructional practices and the roles of learners across education (Han, 2025; Kim, 2025). In particular, generative artificial intelligence (hereafter, “generative AI”) has attracted attention as a transformative educational tool capable of supporting learners’ thinking processes and expanding learning experiences, as it can produce outputs in diverse forms such as text, code, and images (Mun, 2024). Unlike conventional digital tools that primarily provide information, generative AI holds distinctive educational potential in that it can engage in the process through which learners externalize their ideas and subsequently examine and revise them.

Chemistry education has long been identified as a domain in which students experience persistent difficulties in conceptual understanding due to its disciplinary nature, which requires explaining macroscopic phenomena based on interactions at the microscopic level (Chittleborough & Treagust, 2007; Gkitzia et al., 2020; Kim et al., 2025; Park et al., 2020). Interpreting observable phenomena from the perspective of invisible particles imposes a high cognitive demand on learners, often leading to misconceptions or fragmented understanding. To mitigate these difficulties, the use of submicroscopic representations, such as particulate models, has been emphasized, and various visualization tools and digital media have been employed as instructional strategies to bridge macroscopic observations and microscopic explanations.

Within this context, numerous studies in chemistry education have explored the use of simulation, modeling, and programming-based approaches to support students’ conceptual understanding (Kim & Kim, 2022; Kim et al., 2019; Lee et al., 2023). In particular, activities in which students visually represent changes at the particulate level or directly construct and manipulate models have been shown to positively influence their understanding of abstract chemical concepts. However, these approaches also entail a limitation in that they impose a certain level of technical burden. Students are often required to learn programming syntax or tool usage to implement models, and in doing so, cognitive resources may be disproportionately allocated to technical implementation rather than conceptual understanding. Consequently, students may fail to reach the stage of adequately examining, revising, or refining their conceptual understanding, with learning often centered on producing “working code” (Kim, 2024; Oh, 2017). This limitation can be especially pronounced in concept domains such as acid–base equilibrium, where multiple variables interact and require sophisticated reasoning at the microscopic level.

Recently, generative AI has emerged as a potential alternative to address these limitations in educational contexts (Chookaew et al., 2024; Kim et al., 2023). Because generative AI can instantly produce code or explanations in response to learners’ prompts, it can lower the technical barriers to implementation and enable learners to allocate more cognitive resources to conceptual validity. In this way, simulation-based activities have the potential to shift from being implementation-focused tasks to learning experiences centered on conceptual examination and revision. However, the mere use of generative AI, in and of itself, does not guarantee learning outcomes. Generative AI can produce outputs regardless of learners’ level of conceptual understanding, and if these outputs are not critically evaluated, there is a risk of reinforcing misconceptions or fostering superficial understanding (Jho, 2023). Therefore, to meaningfully integrate generative AI into education, instructional designs that encourage learners to examine, compare, and revise AI-generated outputs are essential.

Accordingly, a growing body of research has begun to explore the educational potential of generative AI in chemistry education contexts. For example, Morbidoni et al. (2026) investigated the use of ChatGPT in chemical problem solving and conceptual explanation, demonstrating its potential as a learning support tool. While this study showed that generative AI can provide useful explanations and examples, it did not empirically examine its effects on students' conceptual understanding in actual classroom settings. Similarly, Erümit & Sarıalioğlu (2025) conducted a comprehensive review of AI-based research in science and chemistry education and reported that AI tools have the potential to positively influence learning achievement and engagement. However, they also noted that empirical studies integrating generative AI into concrete instructional activities within chemistry classrooms remain limited. Although these studies highlight the educational potential of generative AI, there is still a lack of empirical research examining how students use generative AI to construct simulations or models and how such activities affect learning outcomes in authentic classroom contexts.

In response, this study aims to examine the educational effects of instruction utilizing generative AI in an authentic classroom setting. This study is distinguished from prior research in that generative AI is not treated merely as a tool for implementing concepts, but as a mediating resource that enables learners to externalize and examine their conceptual understanding. In particular, the instruction was designed around activities in which students directly construct particulate-level simulations using generative AI, compare the generated outputs with chemical concepts, and iteratively revise them. Students used generative AI to create simulations, but rather than accepting the outputs at face value, they repeatedly evaluated their alignment with chemical concepts and corrected or refined errors. This approach holds educational potential in that it reduces the technical burden of implementation while promoting the explicit articulation and reconstruction of learners' conceptual understanding.

This study investigates the educational effects of generative AI–based simulation construction activities from two perspectives: academic achievement and core competencies related to convergence education. Academic achievement serves as a direct indicator of changes in students' understanding of key concepts related to acid–base equilibrium, whereas convergence competencies capture changes in abilities such as creative problem solving, collaboration, and communication, which are required in the process of constructing, revising, and refining simulations. Programming-based modeling and simulation activities have been characterized as complex learning processes involving problem solving, collaborative interaction, and communication. In this regard, prior studies examining the effects of digital-based instruction have analyzed not only academic achievement but also changes in related competencies (Kim & Kim, 2022; Scherer et al., 2019; Yakman & Lee, 2012). Accordingly, this study seeks to comprehensively explore the educational effects of generative AI–supported instruction from the perspectives of conceptual understanding and the development of key competencies. The specific research questions of this study are as follows:

First, how does participation in generative AI–based simulation construction activities on acid–base equilibrium affect students' academic achievement?

Second, how does participation in generative AI–based simulation construction activities on acid–base equilibrium affect students' convergence competencies?

Methods

Participants

This study was conducted with 54 second-year students enrolled in an Advanced Chemistry course during the second semester of 2025 at G Science High School, located in Gyeongsangbuk-do, South Korea. Among the participants, 42 were male (77.8%) and 12 were female (22.2%). Owing to the specialized curriculum structure of science high schools, these students followed a different academic track from those in general high schools. In the first semester of Grade 10, students completed 6 credits of Integrated Science and 3 credits of Chemistry I, while Chemistry II was not included in the curriculum. Instead, an Advanced Chemistry course was offered over three consecutive semesters—from the second semester of Grade 10 to the second semester of Grade 11—with 3 credits per semester, totaling 9 credits. At the time of this study, participants had only learned acid–base concepts through Chemistry I, and the more advanced acid–base concepts addressed in this study were introduced for the first time in the Advanced Chemistry course.

In addition, students completed a total of 6 credits in information-related subjects (3 credits each in the first and second semesters of Grade 10), followed by Information Science courses in Grade 11 (2 credits per semester, totaling 4 credits). Based on this curricular background, participants were considered to possess the foundational digital literacy required to engage in the generative AI–based inquiry activities implemented in this study. Indeed, during instruction, students did not experience significant difficulty in carrying out simulation construction tasks using generative AI.

Prior to the study, the purpose and procedures were explained to both students and their parents, and participation was based on voluntary informed consent. Data were collected and analyzed only from students who agreed to participate, and participants were informed that they could withdraw from the study at any time. Ultimately, data from all 54 students were included in the analysis.

Instructional Procedure

To explore the educational effects of simulation construction using generative AI, this study employed VPython, a Python-based visualization environment, for simulation development. VPython enables the representation of particle motion related to acid–base equilibrium as vector-based dynamic changes and supports three-dimensional visualization. It was therefore selected as an appropriate tool for facilitating learners' intuitive understanding of reaction processes and dynamic changes. The generative AI tool used in this study was ChatGPT, a commercially available system familiar to students, which was utilized primarily to support code generation for simulation implementation.

The instructional procedure was designed with reference to the study by Kim et al. (2019), which examined instructional effects related to acid–base equilibrium. In their study, students used Scratch to directly assign code to sprite objects representing particles, thereby implementing particle motion. In contrast, the present study placed less emphasis on the act of directly coding particle behavior and instead focused on students' conceptual understanding required to generate such code. Students were guided to iteratively examine which aspects of their chemical understanding were inaccurate based on the simulation outputs and to revise their ideas accordingly.

The overall instructional procedure is presented in Table 1. The intervention consisted of 10 class sessions. Excluding the pretest and study orientation conducted in Session 1 and the posttest in Session 10, the remaining eight sessions were devoted to generative AI–based simulation construction activities related to acid–base equilibrium. In Session 2, students were introduced to ChatGPT and VPython, and informed that the target simulation would model the titration of a weak acid, acetic acid (CH_3COOH), with a strong base, sodium hydroxide (NaOH). Rather than addressing individual acid–base concepts in isolation, the weak acid–strong base titration context was selected as the central task because it requires the integrated application of multiple concepts. This context was considered particularly suitable for revealing students' conceptual understanding, as it involves the ionization of a weak acid, neutralization reactions upon base addition, and the subsequent equilibrium processes that determine pH. The titration process inherently links multiple concepts in a sequential and integrated manner.

Simulation construction activities were conducted in small groups. Teams were deliberately composed to include students with varying levels of academic achievement, preventing the concentration of similar ability levels within a single group. During Sessions 3–4, students designed and implemented their initial simulations. In Session 5, feedback was provided on the first set of outputs. During Sessions 6–7, students revised and refined their simulations based on this feedback, followed by additional feedback in Session 8. Finally, students completed their simulations in Session 9.

The instructional design emphasized not merely the construction of simulations, but an iterative process in which students critically examined and revised the code and outputs generated by AI. Working collaboratively, students analyzed the given problem, generated simulations using generative AI, and repeatedly evaluated whether the results were consistent with chemical concepts, making revisions as needed. This process extended beyond conceptual understanding to include exploration and refinement of problem-solving strategies, peer discussion, and collaborative decision-making. However, it should be noted that, due to the collaborative nature of the instruction, there are limitations in interpreting individual academic achievement outcomes, as it is difficult to fully disentangle how understanding developed through group interactions was reflected in individual performance.

The instruction was designed to progress from relatively simple conceptual situations to more complex, integrated ones. In the initial stages, students constructed models focusing on single-equilibrium systems, such as the ionization of acetic acid. Subsequently, the tasks were extended to include hydrolysis of salts and neutralization titration, requiring the integration of multiple equilibrium concepts. Throughout this process, students evaluated whether their conceptual understanding was appropriately represented in the simulations and refined their models by modifying prompts as necessary. Feedback during instruction focused on the alignment between students' prompts, simulation outputs, and underlying chemical concepts. In particular, three key aspects were emphasized: (1) the representation of equilibrium constants (K_a or K_b), (2) the mechanism of interaction in neutralization reactions, and (3) the incorporation of acid–base equilibrium in determining pH. Rather than providing explicit correction answers, feedback was delivered in a question-driven manner to encourage students to independently evaluate and revise their simulations.

Table 1. Instructional procedure of the generative AI–based simulation construction activity

Session	Instructional focus	Activities
1	Pretest and introduction	• Introduction of program goals and collection of consent forms • Administration of academic achievement pretest • Administration of STEAM competencies pretest • Overview of the activity and schedule • Introduction of the titration context (weak acid–strong base: acetic acid–NaOH)
2	Introduction to ChatGPT and VPython	• Introduction to ChatGPT and VPython • Reintroduction of the target task (acetic acid–NaOH titration simulation) • Team formation
3-4	Simulation construction	• Modeling of the titration situation • Structuring concepts of dissociation, neutralization, and dynamic equilibrium • Designing and revising basic code using ChatGPT • Implementing visualization with VPython • Representing particle behavior and pH changes at different NaOH volumes (initial, half-equivalence, equivalence, and 1.5 equivalence points)
5	Feedback	• Review of the first simulation products by teams • Feedback focusing on conceptual and coding errors • Suggestions for revision
6-7	Simulation revision	• Revising variables and structures based on feedback • Refining titration curves and particle representations • Correcting errors and validating the model • Adjusting parameters (Ka, concentration, volume, etc.)
8	Second feedback	• Review of the second simulation products by teams • Feedback focusing on conceptual and coding errors • Guidance for final revision
9	Finalization	• Completion and submission of the final simulation
10	Posttest	• Administration of academic achievement posttest • Administration of STEAM competencies posttest

Survey Tools

The instrument used to measure academic achievement was adapted from the assessment developed by Kim et al. (2019). The original instrument was designed to analyze students' conceptual understanding of acids and bases by presenting four scenarios: ionization of a strong acid, ionization of a weak acid, ionization of a weak base, and the leveling effect. In each scenario, students were required to represent the equilibrium state using an ionic model and explain its meaning.

In the present study, the instrument was revised and refined through discussions involving two doctoral-level researchers in chemical education and one in-service teacher. As a result, the assessment was restructured into five scenarios, as shown in Table 2. Although the instructional intervention in this study focused on the context of weak acid–strong base titration, this context represents an integrated framework that includes the ionization of a weak acid prior to titration, neutralization reactions upon the addition of a base, and the equilibrium and pH interpretation of the resulting species. Accordingly, the items in Table 2 were designed to assess students' understanding of individual concepts underlying the instructional activities. In particular, Items 2 and 5 were directly aligned with the core instructional activities, while Item 4 was designed to indirectly assess students' understanding of the interpretation of species formed after titration. Meanwhile, Items 1 and 3 addressed the ionization of a strong acid and a strong base, respectively, and were included as baseline items to examine students' understanding of the distinction between complete and partial ionization.

Table 2. Contents of the academic achievement measurement instrument

Item No.	Content
1	Ionization of strong acid
2	Ionization of weak acid
3	Ionization of strong base
4	Ionization of weak base
5	Titration of a weak acid with a strong base

Previous research has shown that, although strong acids and strong bases undergo nearly complete ionization in aqueous solution and are thus equivalent in terms of the extent of ionization, students' reasoning and explanatory approaches may vary depending on the specific substances involved (Cooper et al., 2016). Reflecting this finding, the present study included HCl, CH_3COOH, NaOH, and NH_3 in Items 1–4, respectively, to more closely examine students' thinking regarding the ionization of acids and bases across different substances. Item 5 addressed a neutralization titration context, requiring students to represent, using a particulate model, the changes that occur when a sodium hydroxide (NaOH) solution is added to an acetic acid (CH_3COOH) solution and to explain the resulting changes in pH. The internal consistency of the assessment instrument was found to be acceptable, with Cronbach's α values of 0.708 for the pretest and 0.803 for the posttest.

To examine the effects on students' convergence competencies, this study employed the Convergence competencies Inventory for elementary, middle, and high school students developed by Choi et al. (2013). This instrument was designed to assess, in a multidimensional manner, the key competencies required for convergence-based problem solving. The items assess students' competencies in convergence, creativity, caring, and communication.

The instrument consists of 21 items, each rated on a 4-point Likert scale ranging from "strongly disagree" (1) to "strongly agree" (4). As shown in Table 3, convergence competencies was measured across four subdomains, each comprising 4 to 7 items. The internal consistency of the instrument was high, with Cronbach's α values of 0.930 for the pretest and 0.947 for the posttest across all items. At the subdomain level, Cronbach's α ranged from 0.786 to 0.868 in the pretest and from 0.825 to 0.911 in the posttest, indicating generally acceptable to high reliability.

Table 3. Reliability of the STEAM competencies instrument

Category	Item No.	Cronbach's α	
		Pretest	Posttest
Convergence	1, 2, 3, 4, 5	0.788	0.856
Creativity	6, 7, 8, 9, 10, 11, 12	0.868	0.911
Caring	13, 14, 15, 16	0.786	0.825
Communication	17, 18, 19, 20, 21	0.853	0.848
Total		0.930	0.947

Data Collection and Analysis

This study was conducted in an authentic classroom setting to explore the educational effects of generative AI–based simulation construction activities. The participants were students enrolled in an Advanced Chemistry course.

Considering the constraints related to curriculum implementation and fairness in classroom assessment, it was not feasible to establish a control group. Therefore, a one-group pretest–posttest design was employed to examine changes before and after the instructional intervention.

For the academic achievement data, each item was scored dichotomously, with 1 point assigned to scientifically correct responses and 0 points otherwise. A total score was then calculated for each student across the five items. Paired t-tests were conducted to examine pre–post differences at both the item level and the total score level. In addition, Cohen's d was calculated to compare effect sizes across items.

The pretest and posttest for academic achievement were administered individually, and measures were taken to prevent the sharing of answers not only during feedback sessions but throughout the entire instructional process. Furthermore, to explore whether the effects of the generative AI–based simulation activities varied according to students' prior achievement levels, participants were categorized into three groups (high, medium, and low) based on their Advanced Chemistry grades from the first semester of Grade 11. Given the non-normality and distributional characteristics of the data across achievement groups, a nonparametric analysis of covariance (Quade's procedure) was conducted. Specifically, posttest scores were regressed on pretest scores to adjust for initial differences, and the resulting residuals were rank-transformed and used as the dependent variable to test group differences. Post hoc comparisons were conducted using the Bonferroni method.

For the convergence competencies data, mean scores were calculated for each student across subdomains and across all items. Paired t-tests were then conducted to examine pre–post differences at the item level, subdomain level, and overall mean score. Cohen's d was also calculated to compare effect sizes across subdomains.

Results

Changes in Academic Achievement

The results of the pretest and posttest on academic achievement are presented in Table 4. Based on the total score across five items, the mean pretest score was 0.93, whereas the mean posttest score increased to 2.89. A paired t-test indicated that the posttest scores were significantly higher than the pretest scores (t =-8.52, p < .001). The effect size was large (Cohen's d = 1.59).

Table 4. Results of paired t-rest for academic achievement

Item No.	Pretest		Posttest		t	Cohen's d
	Mean	SD	Mean	SD		
1	0.39	0.49	0.97	0.19	-8.45^{***}	1.56
2	0.27	0.46	0.96	0.19	-10.30^{***}	1.96
3	0.09	0.29	0.30	0.46	-2.84^{**}	0.55
4	0.13	0.34	0.28	0.45	-1.83^{*}	0.38
5	0.02	0.14	0.39	0.49	-5.58^{***}	1.03
Total	0.93	1.03	2.89	1.40	-8.52^{***}	1.59

$^{*}p$<.05, $^{**}p$<.01, $^{***}p$<.001

At the item level, all five items showed statistically significant increases from pretest to posttest. The largest effect sizes were observed in the ionization of a strong acid (Item 1, $d = 1.56$) and a weak acid (Item 2, $d = 1.96$). The neutralization titration item (Item 5), which had the lowest pretest mean, also showed a substantial increase ($d = 1.03$). In contrast, the ionization of a strong base (Item 3, $d = 0.55$) and a weak base (Item 4, $d = 0.38$) showed relatively smaller effect sizes compared to the other items.

In the pretest, the highest mean score was observed for the strong acid ionization item (Item 1), whereas the lowest mean score was observed for the weak acid–strong base titration item (Item 5). Although complete ionization was expected to be relatively straightforward, the mean score for the strong base ionization item (Item 3) was comparatively low. Similarly, students demonstrated difficulty not only with the ionization of a weak acid (CH_3COOH, Item 2) but also with the situation involving ammonia (NH_3, Item 4).

Two representative cases illustrate changes in students' conceptual understanding during the simulation construction process. In Team A, students initially assumed that acetic acid completely dissociates in aqueous solution and developed an initial simulation reflecting complete ionization prior to the addition of NaOH (Fig. 1A). Through iterative feedback and revision, the final simulation incorporated the concept of partial dissociation of a weak acid (Fig. 1B).

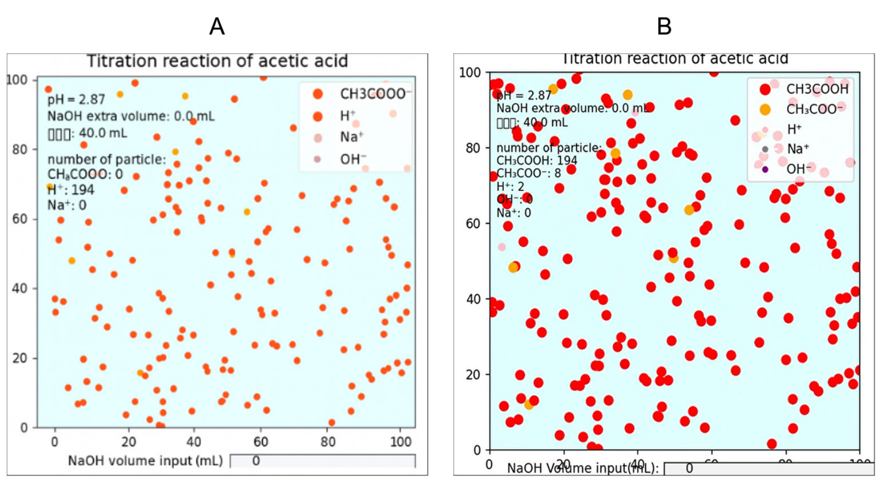

Fig. 1. An example of team A's constructed simulation (A: initial product, B: final product)

In Team B, the initial simulation reflected the misconception that the pH at the equivalence point is 7 (Fig. 2A). After revising their prompts and incorporating feedback, the final simulation included the hydrolysis of CH_3COO^-, resulting in a corrected representation of pH at the equivalence point (Fig. 2B).

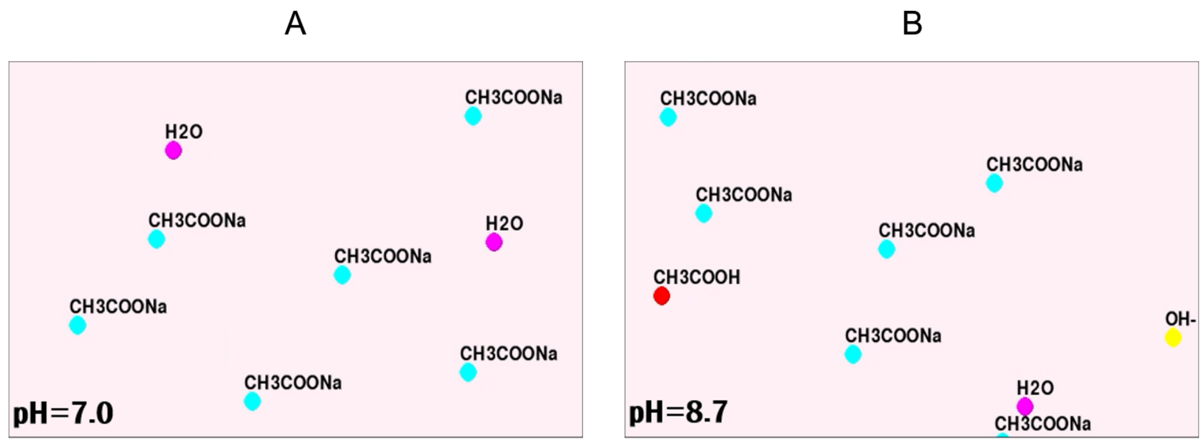

Fig. 2. An example of team B's constructed simulation (A: initial product, B: final product)

To examine whether changes in academic achievement varied by prior achievement level, a nonparametric analysis of covariance (Quade's procedure) was conducted. The results are presented in Table 5. A statistically significant difference was found among the three groups ($p < .001$). The mean ranks of the adjusted residuals were highest in the high-achievement group (MR = 49.33), followed by the medium-achievement group (MR = 31.52) and the low-achievement group (MR = 12.65). Post hoc comparisons using the Bonferroni method indicated that the high-achievement group scored significantly higher than both the medium- and low-achievement groups, and the medium-achievement group scored significantly higher than the low-achievement group ($p < .05$). The effect size for group differences was large ($\eta^2 = .71$).

Table 5. Results of rank-based ANCOVA (Quade procedure)

	SS	df	MS	*F*	Post hoc
Group	9104.71	2	4552.34	62.75***	H>M,L M>L
Error	3699.79	51	72.55		
Total	12804.50	53			

***p<.001

Changes in STEAM Competencies

The results of the pretest and posttest for convergence competencies are presented in Table 6. Based on the overall mean score across the four subdomains, the mean pretest score was 3.40, which increased to 3.62 in the posttest. A paired t-test indicated that the posttest mean was significantly higher than the pretest mean ($p < .001$). The overall effect size was large (Cohen's $d = 0.99$).

Table 6. Results of paired t-test for STEAM competencies

Category	No.	Pretest		Posttest		t	Cohen's d
		Mean	SD	Mean	SD		
Convergence	1	3.76	.47	3.83	.38	-1.27	-
	2	3.65	.56	3.81	.44	-2.63*	0.36
	3	3.59	.57	3.81	.39	-3.04**	0.41
	4	3.54	.54	3.74	.48	-3.05**	0.41
	5	3.63	.53	3.81	.39	-2.63*	0.36
	Sub-total	3.64	.39	3.80	.33	-3.90***	0.53
Creativity	6	2.96	.73	3.26	.68	-3.44**	0.47
	7	3.02	.81	3.22	.74	-2.39*	0.33
	8	3.28	.66	3.44	.63	-2.27*	0.31
	9	3.15	.79	3.31	.77	-2.02*	0.27
	10	3.26	.65	3.56	.63	-4.35***	0.59
	11	3.11	.80	3.56	.60	-4.72***	0.64
	12	3.44	.60	3.78	.42	-4.46***	0.61
	Sub-total	3.14	.54	3.43	.16	-8.69***	1.18
Caring	13	3.37	.78	3.65	.62	-2.98**	0.41
	14	3.28	.76	3.59	.69	-4.01***	0.55
	15	3.74	.48	3.80	.41	-.685	-
	16	3.52	.54	3.67	.48	-1.93	0.26
	Sub-total	3.48	.51	3.67	.45	-3.57**	0.49
Communication	17	3.31	.70	3.61	.66	-3.44**	0.47
	18	3.43	.66	3.65	.56	-2.86**	0.39
	19	3.52	.57	3.72	.45	-3.05**	0.41
	20	3.33	.64	3.54	.69	-1.90	-
	21	3.33	.70	3.59	.57	-3.24**	0.44
	Sub-total	3.39	.52	3.62	.47	-4.44***	0.60
Total		3.40	.42	3.62	.40	-7.27***	0.99

*p<.05, **p<.01, ***p<.001

At the subdomain level, all four areas—convergence, creativity, caring, and communication—showed increases in mean scores from pretest to posttest, and these changes were statistically significant. The effect sizes (Cohen's d) for each subdomain were 0.53 (convergence), 1.18 (creativity), 0.49 (caring), and 0.60 (communication). Among these, creativity showed the largest effect size, whereas the other three subdomains demonstrated moderate effect sizes.

Item-level analyses further showed that all items related to creativity exhibited statistically significant increases from pretest to posttest. In contrast, in the caring and communication subdomains, some items did not show statistically significant changes despite overall increases at the subdomain level.

In the pretest, the highest mean score was observed in convergence (M = 3.64), followed by caring, communication, and creativity, with creativity showing the lowest initial level. In the posttest, mean scores increased across all four subdomains, with creativity showing the most substantial improvement.

Discussions

Conceptual Change in Acid–Base Understanding

The findings of this study suggest that generative AI–based simulation construction activities are associated with meaningful changes in students' conceptual understanding of acid–base equilibrium. The overall effect size observed in this study (d = 1.59) was notably large. In comparison, Kim et al. (2019) reported a large effect size (d = 0.75) in a study employing a similar assessment instrument and a one-group pretest–posttest design using Scratch-based modeling activities. Although a cautious comparison is possible due to similarities in research design and assessment tools, differences in participant characteristics—particularly the inclusion of science high school students in the present study—should be considered when interpreting these results.

The relatively large effect size observed in this study may be attributed, in part, to the alignment between instructional activities and assessment items, as well as the low baseline scores in the pretest. Therefore, the observed effect size should not be interpreted as a generalizable measure of instructional effectiveness but rather as reflecting context-specific learning changes. In addition, the possibility of testing effects due to repeated measures cannot be excluded. More importantly, the research design itself should be considered when interpreting the findings. The one-group pretest–posttest design employed in this study is appropriate for exploring changes within an authentic classroom context; however, it imposes limitations on interpreting the observed changes as causal effects of the instructional intervention.

A key distinction of this study lies in the role of generative AI as a mediating tool for conceptual reflection rather than merely a means of technical implementation. In contrast to prior studies such as Kim et al. (2019), in which students directly assigned code to represent particle behavior, students in this study used generative AI to externalize their conceptual understanding in the form of code and iteratively examined and revised the resulting simulations. This process suggests that generative AI can function as a tool that supports conceptual evaluation and reconstruction, providing a meaningful extension of prior modeling-based approaches in chemistry education.

The item-level results indicate that learning gains were not uniform across concepts. Larger effects were observed in items related to acid ionization, whereas relatively smaller effects were found in base-related items. In particular, students showed greater difficulty with the ionization of NaOH compared to HCl, despite both involving complete ionization. This finding is consistent with prior research by Cooper et al. (2016), which reported that students' reasoning may vary depending on the specific substances involved, even when the underlying chemical principles are equivalent.

Furthermore, Item 4 (NH_3) remained the most challenging even in the posttest. This item requires consideration of proton transfer, in which water acts as a Brønsted–Lowry acid, imposing a higher cognitive demand than simple ionization scenarios. Interestingly, this item appeared more difficult than the neutralization titration context, which is typically regarded as more complex. This result highlights the importance of explicitly addressing the dual role of water as both an acid and a base in instructional design.

The observed differences across items may also be influenced by the instructional context. As the learning activities in this study were centered on weak acid–strong base titration, concepts such as weak acid ionization and neutralization reactions were repeatedly emphasized during simulation construction and feedback processes. In contrast, weak base ionization received relatively less emphasis. Accordingly, some assessment items were closely aligned with the central instructional context, and improvements in those items may reflect not only gains in conceptual understanding but also increased familiarity with the task context. Therefore, differences in learning gains across items should be interpreted with caution, as they may be influenced by both conceptual difficulty and variations in learning opportunities.

The analysis by prior achievement level further suggests that the effects of generative AI–based simulation activities may interact with learners' existing levels of conceptual understanding. The greater gains observed in the high-achievement group indicate that students with stronger prior knowledge may benefit more from such activities. However, given that all participants were drawn from a science high school, these findings should be interpreted as reflecting relative differences within a high-achieving population. Further research is needed to examine whether similar patterns emerge in more diverse educational settings.

Overall, the results indicate that generative AI–based simulation construction activities may promote conceptual change in ways that are sensitive to both the nature of the target concepts and learners' prior knowledge. At the same time, the findings highlight the importance of carefully considering research design and instructional alignment when interpreting the educational effects of such interventions.

Development of Convergence Competencies

The results indicate that generative AI–based simulation construction activities are associated with positive changes in students' convergence competencies. The overall effect size observed in this study (d = 0.99) is considered large and is comparable to or slightly higher than those reported in prior studies using the same instrument. For example, studies applying maker education (d = 0.89) and integrating physical computing into science inquiry (d = 0.93) reported similarly large effect sizes (Choi et al., 2013; Kim et al., 2019). While such comparisons suggest that generative AI–based activities may be effective in promoting convergence competencies, these findings should be interpreted with caution due to the influence of the one-group design and instructional context.

Among the four subdomains, creativity showed the largest effect size (d = 1.18), indicating that the simulation construction activities may have particularly supported students' creative problem-solving abilities. Compared to prior studies (Kim et al., 2019; Lee & Kim, 2020) reporting effect sizes ranging from 0.50 to 0.94 for creativity-related outcomes, the magnitude observed in this study is relatively high. This suggests that the iterative process of generating, evaluating, and refining simulations using generative AI may provide a learning environment conducive to creative problem solving.

In addition, the caring and communication subdomains demonstrated moderate effect sizes. Notably, previous studies have reported relatively small effects in caring (e.g., d = 0.31–0.32; Kim et al., 2019; Lee & Kim, 2020) and limited or non-significant changes in communication (Choi et al., 2013; Jeong & Lee, 2020). In comparison, the present study showed moderate effects in both caring (d = 0.49) and communication (d = 0.60), suggesting that the instructional design may have supported these competencies to a greater extent than in prior interventions.

These changes may be attributed not only to the use of generative AI itself but also to the collaborative and iterative nature of the simulation construction activities. In this study, students worked in teams to design simulations, examine

AI-generated outputs, and revise them through repeated feedback cycles. During this process, students engaged in discussions to interpret discrepancies between expected and simulated results, articulated their ideas, and evaluated their peers' perspectives. Such interactions likely contributed to the development of caring and communication, as students were required to consider others' viewpoints and engage in collaborative decision-making.

For example, when simulation results did not align with expectations, students discussed possible causes within their teams, compared their interpretations, and negotiated shared explanations. Through this process, students externalized their ideas, critically examined alternative perspectives, and collaboratively refined their models. The teacher also facilitated an environment in which students could respect differing opinions and engage in constructive dialogue. These features of the instructional design suggest that convergence competencies such as caring and communication may emerge through structured collaborative interactions embedded in the learning process.

These findings imply that, in designing instruction using generative AI, it is important to consider not only the technological affordances of AI but also the structure of collaborative interaction and feedback processes. Such design elements may play a critical role in supporting the development of convergence competencies.

However, several limitations should be noted. The instrument used in this study was developed to assess general competencies in convergence-based problem solving, and therefore may not fully capture context-specific competencies emerging from domain-specific learning activities. Accordingly, the observed changes in convergence competencies should be interpreted as exploratory evidence of potential impacts associated with generative AI–based simulation activities. Further research is needed to develop more context-sensitive assessment tools and to examine these effects in diverse educational settings.

Conclusions and Implications

This study explored the educational potential of generative AI–based simulation construction activities in the context of acid–base equilibrium learning. By engaging students in the process of generating, evaluating, and refining simulations using generative AI, the instructional design emphasized not only the representation of chemical concepts but also the iterative examination and reconstruction of learners' understanding.

The findings of this study highlight that generative AI can function as more than a tool for automating technical tasks; rather, it can serve as a mediating resource that supports the externalization and refinement of conceptual understanding. In particular, the integration of AI-generated outputs with student-led evaluation and revision processes suggests a shift from implementation-centered activities toward concept-centered learning experiences. This perspective extends prior modeling-based approaches by positioning generative AI as a cognitive and epistemic support tool in chemistry learning.

In addition, the results suggest that the effectiveness of generative AI–based activities may not be uniform across learners or concepts. Instead, its educational impact appears to be shaped by the interaction between learners' prior knowledge, the nature of the target concepts, and the structure of instructional activities. This implies that generative AI may be particularly meaningful when used to support the refinement and reorganization of existing knowledge rather than the initial introduction of new concepts.

From an instructional perspective, this study underscores the importance of designing learning environments that integrate generative AI with collaborative interaction and structured feedback. The findings suggest that combining AI-supported modeling activities with opportunities for discussion, comparison, and revision may enhance not only conceptual understanding but also broader competencies such as creativity, caring, and communication. Thus, the educational value of generative AI lies not solely in its technological capabilities but in how it is pedagogically orchestrated within the learning process.

Despite these implications, several limitations should be acknowledged. The use of a one-group pretest–posttest design limits the ability to draw causal inferences, and the findings may reflect characteristics specific to the high-achieving student population in a science high school. In addition, the alignment between instructional activities and assessment tasks, as well as potential familiarity with the task context, may have influenced the observed changes. Therefore, the results of this study should be interpreted as exploratory evidence of the potential of generative AI–based instruction.

Based on these findings, several directions for future research are suggested. First, further studies employing more rigorous experimental designs, including control groups, are needed to more precisely examine the effects of generative AI–based instruction. Second, research should investigate how the effectiveness of generative AI varies across different stages of learning, such as concept introduction versus conceptual refinement. Third, process-oriented analyses are needed to better understand how students' conceptual understanding evolves during interactions with generative AI, including changes in prompts, interpretations of simulation outputs, and the identification and correction of misconceptions. Finally, future research should extend this approach to a wider range of chemistry topics that involve the interaction between microscopic processes and macroscopic phenomena.

Acknowledgment

This research was supported by the Mid-Career Researcher Program in the Humanities and Social Sciences through the National Research Foundation of Korea(NRF), funded by the Ministry of Education (No. NRF-2026S1A5A2A01003494).

Conflicts of Interest

The author(s) declared no conflicts of interest.

References

Chittleborough, G., & Treagust, D. F. (2007). The modelling ability of non-major chemistry students and their understanding of the sub-microscopic level. Chemistry Education Research and Practice, 8(3), 274–292. https://doi.org/10.1039/B6RP90035F

Choi, Y. H., Noh, J. A., Lim, Y. J., Lee, D. W., Lee, E. S., & Noh, J. H. (2013). The development of the STEAM literacy measurement instrument for elementary, junior-high, and high school students. The Korean Journal of Technology Education, 13(2), 177–198.

Chookaew, S., Kitcharoen, P., Howimanporn, S., & Panjaburee, P. (2024). Fostering student competencies and perceptions through artificial intelligence of things educational platform. Computers and Education: Artificial Intelligence, 7, Article 100308. https://doi.org/10.1016/j.caeai.2024.100308

Cooper, M. M., Kouyoumdjian, H., & Underwood, S. M. (2016). Investigating students' reasoning about acid–base reactions. Journal of Chemical Education, 93(10), 1703–1712. https://doi.org/10.1021/acs.jchemed.6b00417

Erümit, A. K., & Sarıalioğlu, R. Ö. (2025). Artificial intelligence in science and chemistry education: a systematic review. Discover Education, 4(1), Article 178. https://doi.org/10.1007/s44217-025-00622-3

Gkitzia, V., Salta, K., & Tzougraki, C. (2020). Students' competence in translating between different types of chemical representations. Chemistry Education Research and Practice, 21(1), 307–330. https://doi.org/10.1039/c8rp00301g

Han, M. J. (2025). A study on innovation proposals for secondary education in response to changing educational paradigms. Convergence Education Review, 11(2), 123–142. https://doi.org/10.69742/cer.2025.11.2.123

Jeong, E. Y., & Lee, Y. J. (2020). The effect of science reading activities class on science core competency. School Science Journal, 14(4), 501–513. https://doi.org/10.15737/ssj.14.4.202011.501

Jho, H. (2023). Understanding of generative artificial intelligence based on textual data and discussion for its application in science education. Journal of the Korean Association for Science Education, 43(3), 307–319. http://dx.doi.org/10.14697/jkase.2023.43.3.307

Kim, J., Kang, D., & Ko, Y. (2023). A study on educative utilization of generative AI: Focusing on ChatGPT utilization. Journal of the Korean Association of Information Education, 27(6), 691–704. http://dx.doi.org/10.14352/jkaie.2023.27.6.691

Kim, J., Lee, J., & Kim, S. (2025). Enhancing the understanding of dynamic equilibrium of liquid evaporation and vapor condensation. Journal of Chemical Education, 102(9), 3910–3918. https://doi.org/10.1021/acs.jchemed.5c00386

Kim, S. (2024). Analysis of chemistry teachers' perceptions of AI utilization in education: Focusing on participants in first-grade teacher qualification level training. Journal of the Korean Association for Science Education, 44(5), 511–518. http://dx.doi.org/10.14697/jkase.2024.44.5.51

Kim, S. (2025). Issues of scope and sequence in the 2022 revised general elective chemistry curriculum. Journal of Science Education, 49(3), 387–398. http://dx.doi.org/10.21796/jse.2025.49.3.387

Kim, S., Choi, H., & Paik, S. H. (2019). Using a systems thinking approach and a Scratch computer program to improve students' understanding of the Brønsted–Lowry acid–base model. Journal of Chemical Education, 96(12), 2926–2936. https://doi.org/10.1021/acs.jchemed.9b00210

Kim, S. I., Kim, J., Kang, S. J., Kim, T., & Yoon, J. (2019). Development and application of middle school students maker education program using Arduino based on design thinking. Journal of the Korean Institute of Industrial Educators, 44(1), 162–189. https://doi.org/10.35140/kiiedu.2019.44.1.162

Kim, S. K., & Kim, H. J. (2022). The effects of chemistry class using computer-based science inquiry program on positive experiences about science, science core competency, and academic achievement. Journal of the Korean Chemical Society, 66(2), 107–123. https://doi.org/10.5012/jkcs.2022.66.2.107

Lee, S. J., Kim, S. K., & Paik, S. H. (2023). The effect of classes using Scratch for quasi-microscopic representation approaches in dynamic equilibrium learning. Journal of the Korean Chemical Society, 67(4), 241–252. https://doi.org/10.5012/jkcs.2023.67.4.241

Lee, Y., & Kim, T. (2020). The effects of a physical computing convergence class on high school science exploration experiment subjects: SW education recognition and convergence literacy. Journal of Korean Association of Computer Education, 23(4), 69–78. https://doi.org/10.32431/kace.2020.23.4.007

Morbidoni, C., Mascitti, A., Sarra, A., Tonucci, L., & Coccia, F. (2026). Exploring the potential of using ChatGPT in chemistry education. Journal of Chemical Education, 103(1), 286–297. https://doi.org/10.1021/acs.jchemed.5c00854

Mun, S. H. (2024). A research of activity related to convergence teaching and learning methods using artificial intelligence in advanced mathematics courses for gifted students. Convergence Education Review, 10(1), 23–49. https://doi.org/10.69742/cer.2024.10.1.23

Oh, M. J. (2017). Non-major students' perceptions of programming education using the Scratch programming language. Journal of Korean Association of Computer Education, 20(1), 1–11. https://doi.org/10.32431/kace.2017.20.1.001

Park, C. Y., Won, J., Kim, S., Choi, H., & Paik, S. H. (2020). Analysis of problems in the submicro representations of acid–base models in chemistry I and II textbooks of the 2009 and 2015 revised curricula. Journal of the Korean Chemical Society, 64(1), 19–29. https://doi.org/10.5012/jkcs.2020.64.1.19

Scherer, R., Siddiq, F., & Tondeur, J. (2019). The technology acceptance model (TAM): A meta-analytic structural equation modeling approach to explaining teachers' adoption of digital technology in education. Computers & Education, 128, 13–35. https://doi.org/10.1016/j.compedu.2018.09.009

Yakman, G., & Lee, H. (2012). Exploring the exemplary STEAM education in the US as a practical educational framework for Korea. Journal of the Korean Association for Science Education, 32(6), 1072–1086.

Authors' Information

Lee, Hwang-Gi: Gyeongsan Science High School, First Author.

Kim, Sungki: Korea Institute for Curriculum and Evaluation, Associate Research Fellow, Co-corresponding Author. https://orcid.org/0000-0002-0393-4533

Paik, Seoung-Hey: Korea National University of Education, Professor, Co-corresponding Author. https://orcid.org/0000-0002-3822-3838