



Are Two Heads Better Than One?: The Effect of Student-AI Collaboration on Students' Learning Task Performance

Jinhee Kim¹ · Sang-Soog Lee²

Accepted: 19 September 2022 / Published online: 14 October 2022
© Association for Educational Communications & Technology 2022

Abstract

A growing number of educators expect that artificial intelligence (AI) will augment students' capacities and rapidly transform the teaching and learning practice. However, there is a lack of convincing evidence on the effects of Student-AI Collaboration (SAC) on a learning task's performance. A critical examination of the effects on students' learning performance is a crucial step in understanding the potential benefits of SAC on learning. Through a repeated measure experiment participated by 20 undergraduate students in South Korea, this study examined the effects of SAC on a public advertisement drawing task. The findings revealed that SAC significantly affects creativity in content, expressivity in expression, and public utility in effectiveness varied depending on students' attitude toward AI or on the level of drawing skill. Implications for the design of educational AI and AI literacy education are discussed.

Keywords Student-AI collaboration · Effect of human-AI collaboration in education · Artificial intelligence in Education · AI literacy · Educational AI design

Introduction

With the rapid progress and acceleration of Artificial Intelligence (AI) technology in decision-making, prediction, knowledge extraction, and logical reasoning that seemed like the exclusive domain of humans, AI has a profound impact on our society. Such progress in AI, in particular, has led to increased interest in Human-AI collaboration (HAC), where the unique strengths of humans and AI act synergistically. For instance, humans work collaboratively with AI for the sake of improving, optimizing, or automating tasks in their personal lives (Cheng, 2019; Parekh, 2017), work lives (Marijan et al., 2019; Yang et al., 2019), governance (Susar & Aquaro, 2019; Valle-Cruz & Sandoval-Almazan, 2018) and creative

performance such as collaborative writing or art (Liu, 2019; Oh et al., 2018).

Along this line, there is a growing interest and demand to understand how AI supports students to better perform their learning tasks and strengthen students' learning experience. Accordingly, previous studies demonstrated that AI can potentially assist students' learning by serving various roles in its interactions with students, such as a personal tutor (Luckin et al., 2016; Peng et al., 2019) analyzing learning process and outcomes, or collaborative peer (Chen et al., 2020; Kanda et al., 2004) interacting with them to augment capacities to execute their learning task, building and maintaining social relationships and facilitate their learning process. Although the presented results of this body of work elicit that AI uses as a direct teaching and learning tool can have positive effects on students' learning, the transferability of guaranteed positive impacts to an actual student-AI collaboration (SAC) context is questionable. A careful exploration of the resulting effects on students' learning performance is an important step in understanding the potential benefits of SAC on learning. The present study, therefore, aimed to verify the effectiveness of SAC on students' learning task performance. In doing so, the study can provide implications for the instruction on learning with AI and the design of educational AI to

✉ Sang-Soog Lee
kate_ss_lee@korea.ac.kr

Jinhee Kim
Jinhee.Kim@liverpool.ac.uk

¹ School of Future Education, Xi'an Jiatong-Liverpool University, Suzhou, People's Republic of China

² Department of Public Administration, Korea University, 145 Anam-ro, Seongbuk-gu, Seoul 02841, South Korea

facilitate better SAC on learning. The study findings can also serve as convincing evidence of SAC effects.

Literature Review

Existing literature presents us with diverse roles that AI can serve in students' learning.

First, AI's biggest promise in education lies in the personalization of learning and learning materials. In this regard, AI is employed as a personal tutor or a teacher, like the Intelligent Tutoring System (ITS), offering diverse examples, help messages and hints, and step-by-step demonstrations available on-demand to execute the learning task (Desmarais & Baker, 2012). Also, ITS corrects students' misconceptions in the task contents, concepts, and knowledge by tracking the problematic steps in the learning process and providing personalized feedback (VanLehn, 2011). However, it is argued that students' interaction with ITS and their learning are depicted as passive along with a programmed learning path that AI has established. Toward this issue, another stream of research offers AI a role of peer. For example, AI serves a role of less competent learner or a help-seeker, making errors and mistakes on purpose for students to play a role of a competent learner and help-givers to correct AI's mistakes and guide AI. Such learning-by-teaching with AI, students' domain knowledge, and self-efficacy can be improved (Chase et al., 2009). On the other hand, some literature demonstrates the role of AI as a group member or a facilitator in the collaborative discussion by (1) facilitating the formation of teams; (2) managing time, (3) encouraging members to participate evenly, and (4) organizing the members' diverse opinions (Kim et al., 2020). In addition, previous studies find an opportunity to use AI as an improviser to stimulate students' ideation during the creative task process (Lin et al., 2020).

Despite the roles of AI and its positive effects on students' learning, it should also be noted that students' differences have been found to affect the way students interact with AI and the effects of their collaboration. For instance, many studies have indicated individual levels of domain-specific knowledge may lead to varied interaction and experience with technology (Choi, 2015; Popovic, 2003). As Kaptelinin (1996) discussed, the ultimate goal of users in utilizing the technology is to address their unmet demands (problems) of specific problem space (domains) and obtain the finest outcomes and experience, not the skillful use of technology itself. Popovic (2003)'s model further describes how novice and expert users interact with technology differently. The model demonstrates that users with a high level of domain-specific knowledge decide what is the most relevant representation needed by reflecting a profound domain-specific knowledge and applying it into the task domain (TD). In contrast, users with a lower level of domain-specific knowledge have an unstable internal

representation of TD, as they are likely to apply general knowledge to TD instead of in-depth domain-specific knowledge, which makes them challenged when deciding which representation is the best to solve the problem. As a consequence, they are prone to go through a series of trials and errors and repeat the same procedures until making the appropriate decision. Reflecting on the presented results, the problem-solving process between humans and technology and that experience cannot be dissociated from the domain that grounds the content of the technology as well as the knowledge of the users.

In addition, the extensive body of research discusses that the individual's attitude toward technology is closely related to an individual's interaction and engagement with AI (Cruz-Benito et al., 2019; Davis, 1989; Sánchez-Prieto et al., 2019). Technology Acceptance Model (TAM), in particular, has found that the users' attitude toward technology significantly influences an individual's behavioral intention to use the technology, which would affect an individual's actual behavior of adopting the new technology (Abdullah & Ward, 2016; Davis, 1989).

The review of existing literature paves the ways for SAC to make plausible results in the process of performing learning tasks, yet the students' perceived effectiveness of working with AI on their learning tasks and how the SAC impacts the task performance are underexplored. In addition, the effects of SAC may vary depending on students' attitudes toward AI and levels of domain knowledge. The current study, therefore, aimed to examine the SAC effects on students' learning task performance through exploring Korean undergraduate students' perceptions and experiences about the collaboration with AI on executing a learning task. The study, particularly, investigated the differences in the SAC effects on a learning task performance amongst groups of students with differing attitudes toward AI and levels of domain skill with one major research question as follows:

Q. What are the group differences in the SAC effects on a learning task performance?

To address the research questions framed, this study conducted a within-subject experiment in which students worked on the public advertisement drawing task in two experimental conditions: In one, the participants completed the task with suggestions from an AI system, and in the other, without it. The AI that collaborated with students in this study is an algorithm-based system rather than off-the-shelf products or services such as an AI speaker and an intelligent robot.

Methodology

Participant

A total of 20 Korean undergraduate students aged between 22 to 25 years took part in this experiment and were

recruited from different universities in urban areas within or near Seoul, the capital city of South Korea. To achieve purposeful sampling, students' drawing skills and their attitudes toward AI were examined. To be specific, we first collected individual students' drawing work before the experiment to be assessed based on a five-point Likert scale ranging from 1 (highly unsatisfied) to 5 (highly satisfied) by 8 experts (2 professors and 2 experts in art education, 2 professors in public advertisement and 2 experts from an advertisement agency).

As descriptive statistics of drawing assessments from the experts distinctly ascertain the score either above four or below two, the participants in the former were categorized in the high level of drawing skill and the latter in the low level of drawing skills. We then conducted a pre-interview with each participant to discern their entrenched attitude toward AI. Based on Wu et al., (2020), students were grouped into two maximum variations: positive (AI is good, beneficial, helpful, intelligent, with goodwill) versus negative (bad, not beneficial, not helpful, silly, with ill will). Taken both assessments of students' level of drawing skill and attitude toward AI, participants were classified into 4 groups consisting of 5 students: A group with (1) the positive attitude toward AI and the high level of drawing skills (PAHD); (2) the positive attitude toward AI but the low level of drawing skills (PALD); (3) the negative attitude toward AI but the high level of drawing skills (NAHD); (4) the negative attitude toward AI and the low level of drawing skills (NALD) (see Table 1).

AI Technology

Among many AI-assisted drawing tools available, this study took advantage of AutoDraw (www.autodraw.com), a compatible web-based tool from any device whereby interactive principles of AI were applied to convert users' inaccurate and rough input sketches into stylized objects, as illustrated in Fig. 1. Furthermore, AutoDraw was suitable to examine their different perspectives when collaborating with AI, for none of the participants had prior experience of using it.

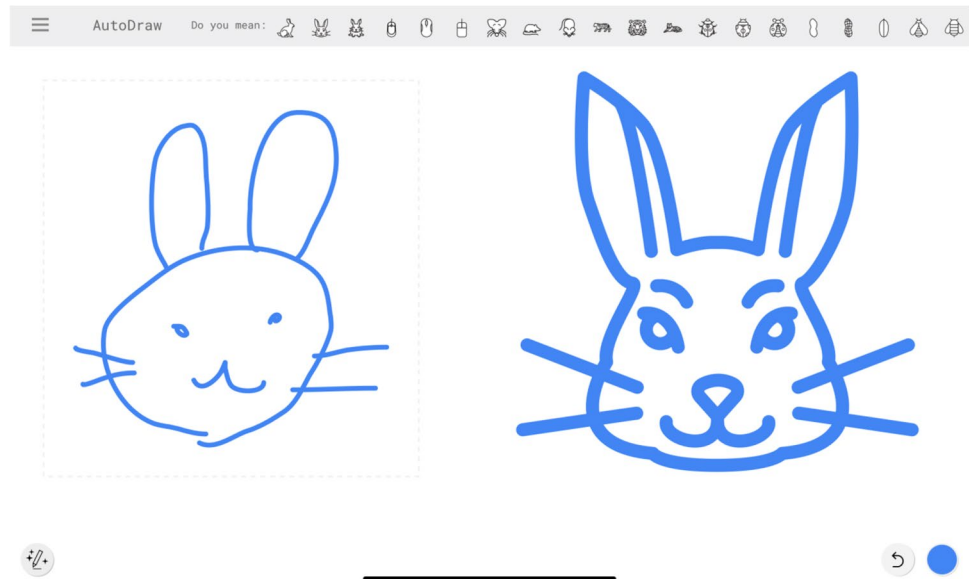
Procedure

The experiment was conducted in four steps. First, the researchers provided a brief explanation about the experiment and the procedure to each participant. Next, participants were asked to watch a short video clip on how to use AutoDraw to learn about major functions. Participants were then given a tablet PC (Galaxy Tab 6 and its smartpen) and opened the AutoDraw Website. Each participant performed two public advertisement drawing tasks: one is to overcome COVID-19, and the other is to cope with climate change. Both were conducted consecutively on the AutoDraw website, while one was conducted in *draw mode* where students worked on the task without AI suggestions, and the other was in *Autodraw mode* where students were given AI suggestions. When they completed it

Table 1 Participants' characteristics and perception of AI

Participant	Gender	Age	Major	Attitude toward AI	Drawing score
PAHD	1 M	24	Technology and Innovation Management	<i>Positive</i>	4.5
	2 F	24	Art education	<i>Positive</i>	4.25
	3 F	22	Nursing	<i>Positive</i>	4.75
	4 F	23	Ceramic arts and design	<i>Positive</i>	4.75
	5 M	24	Industrial design	<i>Positive</i>	4.37
PALD	1 F	22	Computer engineering	<i>Positive</i>	1.75
	2 M	24	Material engineering	<i>Positive</i>	1.62
	3 M	25	Mass Media and Journalism	<i>Positive</i>	1.37
	4 F	23	Business management	<i>Positive</i>	1.62
	5 F	22	Clinical pathology	<i>Positive</i>	1.87
NAHD	1 F	22	Western painting	<i>Negative</i>	4.25
	2 M	24	Visual design	<i>Negative</i>	4.12
	3 F	25	Oriental painting	<i>Negative</i>	4.25
	4 M	24	Film animation	<i>Negative</i>	4.5
	5 F	23	Graphic design	<i>Negative</i>	4.37
NALD	1 F	22	Environmental engineering	<i>Negative</i>	1.87
	2 F	23	International studies	<i>Negative</i>	1.75
	3 F	24	Mass Media and Journalism	<i>Negative</i>	1.37
	4 M	22	Russian language and Literature	<i>Negative</i>	1.5
	5 M	23	Textile engineering	<i>Negative</i>	1.37

Fig. 1 Example of AutoDraw operation



all, a follow-up interview with each participant was carried out for about an hour.

Data Collection and Analysis

We conducted expert evaluations to quantitatively assess the participants' drawing task performance. The evaluation rubric to assess the quality of students' drawing task was first developed based on the literature review (Rourke & Anderson, 2004) that includes three dimensions: Creativity in content, expressivity in expression, and public utility in effectiveness (Eisner, 2002; Pavlou et al., 2000; Stokes, 2005; Stuhlfaut & Yoo, 2013; West et al., 2008) followed by two rounds of expert review administered electronically via email to guarantee the validity of this rubric. Six experts who were proficient in art education, educational

technology, and public advertisement participated in each round of the survey consisting of open-ended questions to verify whether the evaluation areas and criteria were appropriately categorized and make further feedback on how they could be improved. Experts' comments and feedback in round one were analyzed and reviewed (i.e., removing duplications and criteria deemed irrelevant to the study) over multiple sessions by members of the research team. The second round was then once again administered electronically via email to review the revised rubric and to make any suggestions or additional comments on the consolidated list of rubric items. As a result, the final rubric consisting of 9 items with a 5-point Likert scale, ranging from highly unsatisfied (1) to highly satisfied (5) was developed (see Table 2). Then, 8 experts (2 professors in art education, 2 researchers in digital art gallery, 2 teachers in art

Table 2 Evaluation rubric

Category	Evaluation criteria	Score				
		1	2	3	4	5
Creativity (Content)	Is the topic/theme of interest and originality?					
	Is the topic/theme telling the story from a new perspective?					
	Can the final work be interpreted in various ways?					
Expressivity (Expression)	Is the topic/theme effectively expressed and delivered?					
	Is the composition of the final work visually harmonized?					
	Is the digital tool used effectively?					
Public utility (Effectiveness)	Does the final work lead to public empathy?					
	Does the final work induce the actual practice of the public?					
	Is the final work suitable for the public education or promotion materials?					

education, 2 experts in public advertisement) evaluated each students' drawing outcomes on the final rubric and the descriptive statistics of the scores on the drawing task outcomes are summarized in Table 3.

Given the sample size ($n = 160$) of this study, as pre and post of SAC drawing tasks of 20 students were evaluated by 8 different experts, a paired sample T-test would be suitable, and yet the significance in Kolmogorov–Smirnov ($p < 0.001$) and Shapiro–Wilk ($p < 0.001$) did not demonstrate the normality as shown in Table 4. Thus, in this study, we conducted nonparametric statistics of Wilcoxon signed-rank test for the repeated measures on a single sample, using SPSS 23.0 to determine if the drawing task performance in SAC is statistically significant compared to that of on their own (Self).

To improve the accuracy of interpretations drawn from the statistical analysis, semi-structured follow-up interviews were conducted to code for the three dimensions central to the evaluation rubric. The interviews were performed on a one-to-one basis with guiding questions (see Appendix Table 6) such as "what do you think about the drawing outcome?", "do you think that SAC improved creativity in the content on your work?" and "what do you think has positively influenced on creativity in the content on your work?".

All interviews were conducted in Korean and audio-recorded, transcribed, and later translated into English. To compare with the statistical data, categorical analysis was used (Coffey & Atkinson, 1996) to identify passages relating to each of the three dimensions of the evaluation rubric. The interview data were coded using a deductive approach in the first cycle, starting based on the three dimensions of

the evaluation rubric (Miles et al., 2020). Subsequently, a total of 5 themes were grouped under the three dimensions to understand and explain the effects of SAC on drawing outcomes. The qualitative data or the embedded components may not be independent of the larger study context but provides additional knowledge linked to the primary aims of the study; therefore, it is critical to the present study (Plano Clark et al., 2013). As a validity check, two external independent researchers reviewed and cross-checked data with the study findings. Also, any disagreement between researchers was discussed and clarified until a consensus was achieved.

Findings

According to the Wilcoxon Signed-Rank Test (see Table 5), by and large, the total scores of all groups are found to demonstrate significant differences. To be specific, when SAC occurs, PALD ($z = -5.06, p < 0.001$) improved their performance by 2.35 points, PAHD ($z = -5.08, p < 0.001$) by 1.60 and NALD ($z = -4.47, p < 0.001$) by 1.28 whereas the score of NAHD ($z = -4.79, p < 0.001$) was rather lowered by 2.17. The differences in total scores of all groups, thus, indicate the collaboration with and without AI do differentiate the final outcomes.

The Effects of SAC on Creativity in Content

In order to examine the detailed differences in categories of the drawing performance, the scores in creativity, expressivity, and public utility were tallied. In light of creativity, the drawing task scores of PAHD ($z = -5.19, p \leq 0.001$) and

Table 3 Descriptive statistics of the scores on the drawing task outcomes

Category	Self			SAC		
	N	Mean	SD	N	Mean	SD
Creativity	160	2.86	1.06	160	3.19	1.25
Expressivity	160	3.04	1.45	160	3.54	1.04
Public utility	160	2.80	1.18	160	2.73	1.33
Total score (15)	160	8.70	3.21	160	9.46	2.82

Table 4 The normality tests on the drawing scores

Category	Self						SAC					
	Kolmogorov–Smirnov			Shapiro–Wilk			Kolmogorov–Smirnov			Shapiro–Wilk		
	Statistic	df	Sig	Statistic	df	Sig	Statistic	df	Sig	Statistic	df	Sig
Creativity	.227	160	.000	.884	160	.000	.186	160	.000	.907	160	.000
Expressivity	.189	160	.000	.877	160	.000	.197	160	.000	.893	160	.000
Public utility	.196	160	.000	.904	160	.000	.176	160	.000	.892	160	.000
Total score	.106	160	.000	.960	160	.000	.124	160	.000	.938	160	.000

Table 5 Wilcoxon signed-rank test results of the SAC effects

Category	Group	Self		SAC		Differences (SAC-Self)		Negative ranks			Positive ranks			Wilcoxon's			
		M (SD)		M (SD)		N		Mean Rank		Rank Sum		Mean Rank		Rank Sum		Z (p)	
		N	Mean Rank	N	Mean Rank	N	Mean Rank	N	Mean Rank	N	Mean Rank	N	Mean Rank	N	Mean Rank	N	Mean Rank
Creativity (Content)	PAHD	3.03 (0.97)	4.18 (0.68)	1.15	0	0.00	0.00	0.00	33	17.00	561.00	-5.190 (.000)					
	PALD	2.23 (0.66)	3.58 (0.81)	1.35	0	0.00	0.00	0.00	37	19.00	703.00	-5.496 (.000)					
	NAHD	3.63 (1.05)	3.30 (1.09)	-0.33	14	9.46	132.50	38.50	4	9.63	38.50	-2.168 (.030)					
	NALD	2.55 (0.99)	1.70 (0.79)	-0.85	31	16.00	496.00	0.00	0	0.00	0.00	-5.353 (.000)					
Expressivity (Expression)	PAHD	4.33 (0.66)	4.35 (0.70)	0.02	1	2.00	2.00	4.00	2	2.00	4.00	-5.577 (.564)					
	PALD	1.88 (0.72)	3.60 (0.87)	1.72	0	0.00	0.00	780.00	39	20.00	780.00	-5.547 (.000)					
	NAHD	4.30 (0.69)	3.00 (1.06)	-1.30	33	17.00	561.00	0.00	0	0.00	0.00	-5.135 (.000)					
	NALD	1.68 (0.69)	3.28 (0.85)	1.60	0	0.00	0.00	820.00	40	20.50	820.00	-5.655 (.000)					
Public Utility (Effectiveness)	PAHD	3.18 (1.24)	3.60 (1.30)	0.42	3	8.00	24.00	166.00	16	10.38	166.00	-3.038 (.002)					
	PALD	2.28 (0.88)	1.55 (1.11)	-0.73	29	17.07	495.00	100.00	5	20.00	100.00	-3.491 (.000)					
	NAHD	3.63 (1.03)	3.10 (1.13)	-0.53	19	13.29	252.50	47.50	5	9.50	47.50	-3.089 (.002)					
	NALD	2.31 (0.82)	2.65 (0.80)	0.34	1	19.50	19.50	211.50	20	10.58	211.50	-3.567 (.000)					
Total	PAHD	10.53 (2.43)	12.13 (2.31)	1.60	0	0.00	0.00	561.00	33	17.00	561.00	-5.081 (.000)					
	PALD	6.38 (1.88)	8.73 (2.29)	2.35	0	0.00	0.00	561.00	33	17.00	561.00	-5.059 (.000)					
	NAHD	11.55 (2.36)	9.38 (2.81)	-2.17	32	18.92	605.50	24.50	3	8.17	24.50	-4.787 (.000)					
	NALD	6.35 (1.99)	7.63 (1.60)	1.28	2	14.75	29.50	498.50	30	16.62	498.50	-4.466 (.000)					

PALD ($z = -5.50, p \leq 0.001$) groups ascended respectively by 1.15 and 1.35 points while NAHD ($z = -2.17, p < 0.05$) and NALD ($z = -5.35, p \leq 0.001$) were lessened by 0.33 and 0.85 accordingly. It is noted that PA groups demonstrate creativity enhancement on their tasks when SAC happens.

During the interviews, there were two major differences found between PA groups (students with the positive attitude toward AI) and NA groups (students with the negative attitude toward AI). First, it was found that PA groups actively engaged in separate phases for problem definition (i.e., activating previous knowledge and experience to address the problem, framing the boundaries of the problem, clarifying the problem through the specification of criteria and constraints of the problem) and problem-solving (i.e., generating potential solutions, clarifying potential solutions through evaluation of the analysis and synthesis to the problem) prior to the collaborative drawing with AI whereas NA groups tended to jump ahead to draw pictures without adequately performing a prior reflection on the problem and outlining solutions.

I didn't really care about how to draw well since I was quite confident in my drawing skills. Moreover, I could even use AI images. So I tried to focus on problem identification and what to draw to address the problem to make an advertisement original and appealing. (PAHD 4)

As there is saying, 'a problem well stated is a problem half solved,' I spent quite a lot of time thinking about what problems to address within the task themes given. I think that is even more important in SAC. You know, AI can help me with the drawing part. So I put more effort into activating my prior experiences and knowledge to frame the problem within the task themes and generate creative solutions to address the problem. (PALD 3)

Second, although both PA and NA groups highlighted the AI's low accuracy that it did not suggest exactly what they expected, PA groups actively create a story by providing meaning to the arrangement of AI suggested figures, reestablishing the conceptual structure, and developing alternative ideas while NA groups were eager to find the images that exactly match with their intention. For instance, PAHD 3 described her experience saying,

There were moments that I wondered about AI's suggestions and felt disappointed about its inability to recognize my drawing, but I tried with my best to add meaning to AI's suggestion, connecting my sketch and AI's figure by developing a story. By doing such data

storytelling, I became more flexible to utilize AI's awkward recommendations.

Her quotes suggest haphazard element placements in the middle of the SAC could be a source of activating students' storytelling skills. This, then, poses us to think about how to extend AI in visualization software development to enable storytelling. More research is needed to develop interfaces that combine visualization construction during the SAC with the specification of narrative structure, textual/graphical annotation, visual highlighting techniques, transitions, and necessary instruction to augment students' creativity during SAC.

The Effects of SAC on Expressivity in Expression

In the aspect of expressivity, All but PAHD revealed statistically significant differences in expression scores on drawing tasks before and after the SAC. The task performances through the SAC of PALD ($z = -5.55, p \leq 0.001$) and NALD ($z = -5.66, p \leq 0.001$) were assessed to be higher by 1.72 and 1.60 points respectively when that of NAHD ($z = -5.14, p \leq 0.001$) fell by 1.3 points. It is noteworthy that the task performances of LD groups (students with the low level of drawing skills) turned out much better when the SAC took place.

During the interview, entire LD groups expressed that their power of expression has been promoted through interaction with AI. Students were most satisfied with and appreciated the AI's ability to perceive their rough scribbles, generate the list of symmetrical icons or clip-art-style pieces to choose from, and convert their original sketch, which enabled them to visualize what they implicitly think. The following quotes well represent these views:

AutoDraw comprehends my drawing of human hands that looked like chicken feet and suggests a list of different hand shapes! (PALD 4)

I know how a turtle looks but it was hard to visualize a turtle that I thought of in my mind. But thanks to AutoDraw, I could draw a turtle which is the main character in this poster. (NALD 2)

However, it should also be noted that students highlighted the importance of an appropriate level of automation (LOA) to provide personalized scaffolding in the drawing task process. Students in this group wanted detailed instruction rather than automatically showing a list of pre-set figures and converting their hand-drawn images. For instance, PALD 2 expressed:

It's good that I have completed the task in good quality in a shorter time. However, I wonder if that's meaningful interaction in terms of learning. I'd rather want AI

to guide me how to draw things step by step so that I can become an independent artist.

Similarly, NALD 1 mentioned,

Compared to the outcome drawn all by myself, the outcome in collaboration with AI is more refined. But it would have been a much more meaningful experience if AutoDraw offered me just-in-time feedback when I struggled to draw the COVID-19 virus shape or provided me necessary instructional support to try different drawing techniques.

Besides, students compared their drawings with those of the AI and felt disappointed with their drawing skills, which demeaned their confidence. For instance, NALD 4, comparing the part the AI had drawn to that of her own, said, "*The part I drew remained as the fly in the ointment. Perhaps, I should have deleted the part I drew.*"

On the other hand, HD groups (students with the high level of drawing skills) were unsatisfied and disappointed with AI's drawing style, assessing its drawing style as childish and simplistic. They sought more advanced drawing techniques with various materials and develop their current drawing skills to another level. Besides, LAHD opposed automatic transformations of their drawings, describing those as a Territorial issue. For example, NAHD 3 said,

AutoDraw should show respect to my effort, knowledge, and experience in drawing. How could it covert my entire sketches into its own? It had been attempting to overcome my drawing territory!

Similarly, NAHD 4 said, "*I just wanted to partially take AutoDraw's suggestion. But it was greedy that it wanted to control over all my drawing.*" Their view indicates that although AI's automation contributes to the task performance enhancement, it should reflect levels of students' domain-specific skills and students' agency (Wang et al., 2019).

The Effects of SAC on Public Utility in Effectiveness

Moreover, there were significant differences in public utility in all groups as well. The positive changes were found in PAHD ($z = -3.04$, $p < 0.01$) by 0.42 and NALD ($z = -3.57$, $p < 0.001$) by 0.34 while the opposite was true in groups of PALD ($z = -3.49$, $p < 0.001$) by 0.73 and NAHD ($z = -3.09$, $p < 0.01$) by 0.53.

During the interviews, both PAHD and NALD reported that their frequent evaluation and task appraisal during the

SAC would positively influence the public utility in effectiveness. For example, PAHD 2 said:

I constantly reminded myself that I was drawing a public advertisement, not a freestyle of drawing. So I regularly checked whether my images and AI-suggested images were in line to deliver effective messages to the public.

On the other hand, PALD and NAHD tended to spend time mostly on monitoring interaction with AI, assessing AI's suggestions, and adjusting the drawing strategy to complete the task. They mainly evaluate the learning task process and content of the task at the end of the SAC.

I think I simply enjoyed interacting with AI. It was just so much fun that it reacted to my sketches and showed me tens of exemplary figures that I can go through and pick one. I should admit that I have forgotten the purpose of drawing. (PALD 5)

This finding is in line with the earlier research in a student–student collaborative learning context, which found adaptive assessment in collaborative learning leads to an improved understanding of task objective and achievement in task process during the collaborative learning process, which in turn positively influence the task performance (Pifarré & Argelagós, 2020).

Discussion and Conclusion

Based on the Wilcoxon signed-rank test results, unlike NAHD, the overall drawing task performances of PAHD, PALD, and NALD enhanced when students collaborated with AI as a team. Considering three categories in tandem with the interviews, the positive improvements on creativity were found solely in PA groups that enthusiastically indulged in discrete phases of problem definition and problem-solving prior to the SAC; the embellishment on expressivity was shown in LD groups with the help of AI's automation; and much more refined public utility was found in PAHD and NALD.

In light of the findings, a series of implications for educational AI design and instructional strategies for the SAC to be better structured to positively impact students' learning task performance. For the design of educational AI that collaborates with students on a learning task, this study first suggests that automation in students' learning context should be regarded as adaptive scaffolding, not an all-or-nothing phenomenon. Although the LD groups' power of expression has significantly improved with the help of

AI's automated figure, they would rather seek scaffolding-driven automation that can offer them detailed instructional support in need. On the other hand, HD groups were disappointed by AI's limited drawing style and sought an advanced level of drawing techniques to improve their drawing skills to a higher level. In educational research, scaffolding implicates providing needs-driven assistance to students, dwindling it away as their competence increases (Hogan & Pressley, 1997). In this regard, flexible and adaptive AI that can vary its LOA in consideration of students' differences such as levels of domain skills, learning strategies employed and learning goals (Ley et al., 2010) during learning tasks to better provide conceptual, metacognitive, procedural and strategic scaffolding to positively enhance learning (Hannafin et al., 1999). Second, it was evident that the effects of the SAC are cross-linked on various learning activities in the learning process. The effects of SAC on creativity in content were particularly found to be different depending on students' engagement in problem definition and solution generation activity prior to the collaborative drawing with AI. In this respect, AI performance should not be limited to data manipulation or automation. Instead, AI should be developed in a way that can systematically measure students' learning progress and augment or supplement students' task performance capacity throughout the learning process. As the previous literature demonstrated, AI could motivate concept exploration, provoke unexpected ideas and engage students in the ideation process (Lin et al., 2020). AI could also assist students to explore diverse instruments and ways to execute the task during the actual times of collaboration.

In parallel to the development of educational AI, it is essential to provide adequate instructional support, particularly by art-related teachers, to enhance SAC for its positive effects on students' learning task performance throughout the SAC process. The finding revealed that the students who achieved better performance in creativity conducted the problem identification and solution generation activities. This finding coincides with the creative problem-solving process (CPS) model that consists of six steps: (1) problem finding, (2) problem definition, (3) solution finding, (4) idea evaluation, (5) implementation, and (6) acceptance finding (i.e., Hogan & Pressley, 1997; Torrance et al., 1978). Previous research has demonstrated the effectiveness of the CPS on enhancing one's creative thinking ability (Lim & Han, 2020; Wang & Horng, 2002), interplaying between one's divergent thinking to produce a large number of ideas and one's convergent thinking to evaluate or judge the value of each idea for further development. In this regard, teachers should allow students to be engaged in the key components of

the CPS, such as problem identification/definition from behind a mess, alternative solution development that might solve the problem, evaluation of these alternatives that are closely related to the cognitive stages in problem-solving, and implementation (i.e., Klahr & Simon, 1999; Newell & Simon, 1972) in the SAC process. Along the way, it was clear that students found storytelling extremely useful to establish conceptual structure, provide meaning on AI's suggested figures, and even develop new images by combining AI suggestions with their own ideas. Previous research had already demonstrated that the process of explanation and storytelling increased critical and creative thinking and understanding by pushing students to explain the consequences of their views and to search for new information needed for answering questions and achieving cognitive goals (Scardamalia & Bereiter, 2014). This implies that teachers should encourage students to talk through the data suggested by AI (data-telling) and create meaning behind data to enrich their creativity during the SAC. Furthermore, it was shown that the students who conducted frequent evaluations and task appraisals during SAC demonstrated better performance on the public utility in effectiveness. This finding is corroborated with the earlier literature highlighting that the constant evaluation during collaboration forced the task performance as individuals re-evaluate the details of the work (Valkenburg & Dorst, 1998). In this respect, it is essential to activate 'reflection-in-action' during the SAC that students actively reappraise the task process, the content of the task, as well as their actions as individuals and as collective teams to reframe their process (McNiff & Whitehead, 2011).

This study contributes to the research on AI in Education (AIED) and HAC in the education context by examining the effects of the SAC on a drawing task. The study, however, has a few limitations that need to be considered in future studies. First, the number of research participants needs to be increased to enhance the statistical power of a test. Second, this study limited the experimental context to a public advertisement drawing task with algorithmic AI, AutoDraw. Thus, future research is necessary to determine whether the SAC improves students' learning task performance with other types of AI (i.e., chatbot, social robot, AI speaker) in other learning tasks. Third, this study focused on the collaboration between an individual student and an AI algorithm. Future studies can consider the SAC in a large group of students (a group) and AI from a macroscopic point of view. For instance, investigating differences between the effects of the SAC at an individual level and group level could be considered.

Appendix

Table 6 Interview guide

Category	Questions
General	What do you think about the drawing outcome? What is the major difference between times when performing a drawing task with AI and times when performing the task on your own? Why do you think such a difference occur?
Creativity (Content)	Do you think that SAC improved creativity in the content on your work? What do you think has positively influenced on creativity in the content on your work?
Expressivity (Expression)	Do you think that SAC improved expressivity in the expression on your work? What do you think has positively influenced on expressivity in the expression on your work?
Public utility (Effectiveness)	Do you think that SAC improved public utility in the effectiveness on your work? What do you think has positively influenced on public utility in the effectiveness on your work?

Declarations

Informed Consent This study received ethical approval from the university's Institutional Review Board and informed consent from all participants.

Research Involving Human Participants and/or Animals 20 Korean undergraduate students participated in the study.

Disclosure of Potential Conflicts of Interest There is no conflict of interest in this study.

References

- Abdullah, F., & Ward, R. (2016). Developing a general extended technology acceptance model for E-learning (GETAMEL) by analysing commonly used external factors. *Computers in Human Behavior*, *56*, 238–256. <https://doi.org/10.1016/j.chb.2015.11.036>
- Chase, C. C., Chin, D. B., Oppezzo, M. A., & Schwartz, D. L. (2009). Teachable agents and the protégé effect: increasing the effort towards learning. *Journal of Science Education and Technology*, *18*(4), 334–352. <https://doi.org/10.1007/s10956-009-9180-4>
- Chen, H., Park, H. W., Zhang, X., & Breazeal, C. (2020, March). Impact of interaction context on the student affect-learning relationship in child-robot interaction. In *Proceedings of the 2020 ACM/IEEE International Conference on Human-Robot Interaction* (pp. 389–397).
- Cheng, T. (2019). Product recommendation system design. In *Proceedings of the 2019 2nd International Conference on Information Management and Management Sciences* (pp. 71–74).
- Choi, J. M. (2015). Effects of users' domain knowledge on user experience: Focused on the UI design of mobile applications. *Journal Korea Society of Visual Design Forum*, *49*, 253–262.
- Coffey, A. J., & Atkinson, P. (1996). *Making sense of qualitative data: Complementary research strategies (and social thought)*. SAGE Publications Inc.
- Cruz-Benito, J., Sánchez-Prieto, J. C., Therón, R., & García-Peñalvo, F. J. (2019). Measuring students' acceptance to AI-driven assessment in eLearning: Proposing a first TAM-based research model. In *International conference on human-computer interaction* (pp. 15–25).
- Davis, F. D. (1989). Perceived usefulness, perceived ease of use, and user acceptance of information technology. *MIS Quarterly*, *13*(3), 319. <https://doi.org/10.2307/249008>
- Desmarais, M. C., & Baker, R. S. J. D. (2012). A review of recent advances in learner and skill modeling in intelligent learning environments. *User Modeling and User-Adapted Interaction*, *22*(1), 9–38. <https://doi.org/10.1007/s11257-011-9106-8>
- Eisner, E. W. (2002). *The arts and the creation of mind*. Yale University Press.
- Hannafin, M., Land, S., & Oliver, K. (1999). Open learning environments: Foundations, methods, and models. *Instructional-Design Theories and Models: A New Paradigm of Instructional Theory*, *2*, 115–140.
- Hogan, K. E., & Pressley, M. E. (1997). *Scaffolding student learning: Instructional approaches and issues*. Brookline Books.
- Kanda, T., Hirano, T., Eaton, D., & Ishiguro, H. (2004). Interactive robots as social partners and peer tutors for children: A field trial. *Human-Computer Interaction*, *19*(1), 61–84. <https://doi.org/10.1080/07370024.2004.9667340>
- Kaptelinin, V. (1996). Activity theory: Implications for human-computer interaction. In B. A. Nardi (Ed.), *Context and consciousness: Activity theory and human-computer interaction* (pp. 64–103). The MIT Press.
- Kim, S., Eun, J., Oh, C., Suh, B., & Lee, J. (2020, April). Bot in the bunch: Facilitating group chat discussion by improving efficiency and participation with a chatbot. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (pp. 1–13).
- Klahr, D., & Simon, H. A. (1999). Studies of scientific discovery: Complementary approaches and convergent findings. *Psychological Bulletin*, *125*(5), 524–543. <https://doi.org/10.1037/0033-2909.125.5.524>
- Ley, T., Kump, B., & Gerdenitsch, C. (2010). Scaffolding self-directed learning with personalized learning goal recommendations. In *International Conference on User Modeling, Adaptation, and Personalization* (pp. 75–86).
- Lim, C., & Han, H. (2020). Development of instructional design strategies for integrating an online support system for creative problem solving into a University course. *Asia Pacific Education Review*, *21*(4), 539–552. <https://doi.org/10.1007/s12564-020-09638-w>
- Lin, Y., Guo, J., Chen, Y., Yao, C., & Ying, F. (2020). It is your turn: Collaborative ideation with a co-creative robot through sketch. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems (CHI '20)*. NY: Association for Computing Machinery.

- Liu, S. (2019). Everybody's Song Making. *Performance Research*, 24(1), 120–128. <https://doi.org/10.1080/13528165.2019.1594267>
- Luckin, R., Holmes, W., Griffiths, M., Corcier, L. B., UCL Knowledge Lab, Pearson (Firm), & University College, L. (2016). *Intelligence Unleashed*. Pearson.
- Marijan, D., Shang, W., & Shukla, R. (2019). Implications of resurgence in artificial intelligence for research collaborations in software engineering. *ACM SIGSOFT Software Engineering Notes*, 44(3), 68–70. <https://doi.org/10.1145/3356773.3356813>
- McNiff, J., & Whitehead, J. (2011). *All you need to know about action research*. Sage Publications.
- Miles, M. B., Huberman, A. M., & Saldaña, J. (2020). *Qualitative data analysis: A methods sourcebook*. Sage Publication.
- Newell, A., & Simon, H. A. (1972). *Human problem solving*. Prentice-Hall.
- Oh, C., Song, J., Choi, J., Kim, S., Lee, S., & Suh, B. (2018). I lead, you help but only with enough details: Understanding user experience of co-creation with artificial intelligence. *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (pp. 1–13).
- Parekh, R. (2017). Designing AI at scale to power everyday life. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 27–27).
- Pavlou, P. A., & Stewart, D. W. (2000). Measuring the Effects and Effectiveness of Interactive Advertising. *Journal of Interactive Advertising*, 1(1), 61–77. <https://doi.org/10.1080/15252019.2000.10722044>
- Peng, H., Ma, S., & Spector, J. M. (2019). Personalized adaptive learning: An emerging pedagogical approach enabled by a smart learning environment. *Smart Learning Environments*, 6(1), 1–14. <https://doi.org/10.1186/s40561-019-0089-y>
- Pifarré, M., & Argelagós, E. (2020). Embedded information problem-solving instruction to foster learning from digital sources: Longitudinal effects on task performance. *Sustainability*, 12(19), 7919. <https://doi.org/10.3390/su12197919>
- Plano Clark, V. L., Schumacher, K., West, C., Edrington, J., Dunn, L. B., Harzstark, A., Melisko, M., Rabow, M. W., Swift, P. S., & Miaskowski, C. (2013). Practices for embedding an interpretive qualitative approach within a randomized clinical trial. *Journal of Mixed Methods Research*, 7(3), 219–242. <https://doi.org/10.1177/1558689812474372>
- Popovic, V. (2003). Expert and novice users model and their application to the design process. *Journal of the Asian Design International Conference*, 1, 1–12.
- Rourke, L., & Anderson, T. (2004). Validity in quantitative content analysis. *Educational Technology Research and Development*, 52(1), 5–18.
- Sánchez-Prieto, J. C., Cruz-Benito, J., Therón, R., & García-Peñalvo, F. J. (2019). How to measure teachers' acceptance of AI-driven assessment in eLearning: A TAM-based proposal. In *Proceedings of the Seventh International Conference on Technological Ecosystems for Enhancing Multiculturality* (pp. 181–186).
- Scardamalia, M., & Bereiter, C. (2014). Cambridge handbook of the learning sciences. In R.K. Sawyer (2nd Ed.), *Knowledge building and knowledge creation: Theory, pedagogy, and technology* (pp. 397–417). Cambridge University Press.
- Stokes, P. D. (2005). *Creativity from constraints: The psychology of breakthrough*. Springer.
- Stuhlfaut, M. W., & Yoo, C. Y. (2013). A tool for evaluating advertising concepts: Desirable characteristics as viewed by creative practitioners. *Journal of Marketing Communications*, 19(2), 81–97. <https://doi.org/10.1080/13527266.2010.550009>
- Susar, D., & Aquaro, V. (2019). Artificial intelligence: Opportunities and challenges for the public sector. *Proceedings of the 12th International Conference on Theory and Practice of Electronic Governance* (pp. 418–426).
- Torrance, E. P., Torrance, J. P. Willams, S. J., Horng, R.-Y., & Crable, A. B. (1978). *Handbook for training future problem solving teams*. Lincoln, Nebraska: Future Problem Solving Bowl.
- Valkenburg, R., & Dorst, K. (1998). The reflective practice of design teams. *Design Studies*, 19(3), 249–271. [https://doi.org/10.1016/s0142-694x\(98\)00011-8](https://doi.org/10.1016/s0142-694x(98)00011-8)
- Valle-Cruz, D., & Sandoval-Almazan, R. (2018). Towards an understanding of artificial intelligence in government. In *Proceedings of the 19th Annual International Conference on Digital Government Research: Governance in the Data Age* (pp. 1–2).
- VanLehn, K. (2011). The Relative effectiveness of human tutoring, intelligent tutoring systems, and other tutoring systems. *Educational Psychologist*, 46(4), 197–221. <https://doi.org/10.1080/00461520.2011.611369>
- Wang, D., Weisz, J. D., Muller, M., Ram, P., Geyer, W., Dugan, C., Tausczik, Y., Samulowitz, H., & Gray, A. (2019). Human-AI collaboration in data science. In *Proceedings of the ACM on Human-Computer Interaction*, 3(CSCW), 1–24. <https://doi.org/10.1145/3359313>
- Wang, C. W., & Horng, R. Y. (2002). The effects of creative problem solving training on creativity, cognitive type and R&D performance. *R&D Management*, 32(1), 35–45. <https://doi.org/10.1111/1467-9310.00237>
- West, D. C., Kover, A. J., & Caruana, A. (2008). Practitioner and customer views of advertising creativity: same concept, different meaning? *Journal of Advertising*, 37(4), 35–46. <https://doi.org/10.2753/joa0091-3367370403>
- Wu, Y., Mou, Y., Li, Z., & Xu, K. (2020). Investigating American and Chinese Subjects' explicit and implicit perceptions of AI-Generated artistic work. *Computers in Human Behavior*, 104, 106186. <https://doi.org/10.1016/j.chb.2019.106186>
- Yang, Q., Steinfeld, A., & Zimmerman, J. (2019). Unremarkable AI: Fitting intelligent decision support into critical, clinical decision-making processes. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (pp. 1–11).

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Springer Nature or its licensor holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.