

Artistic Practice Opportunities in CST Evaluations: A Longitudinal Group Deployment of ArtKrit

ANONYMOUS FOR SUBMISSION

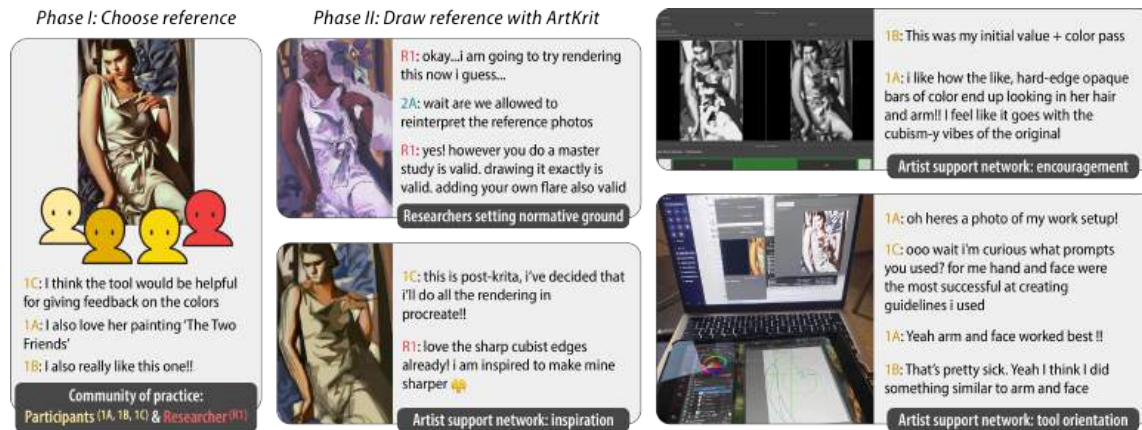


Fig. 1. A sample of one group's week during our three-week evaluation. Nine participants were divided into three small communities of practice that included a researcher; each week, the group would vote on and draw a "master study". Each group had their own Discord channel, visible to all participants. Participant comments are real Discord messages sent. Shown are works-in-progress shared from Week 1 of Group 1; Figure 4 shows final results. Our study structure encouraged building artist support networks [13] and observing a mini creativity support ecosystem [1]. Reference image: Portrait of Madame M. by Tamara De Lempicka.

Creativity support tools (CSTs) aim to elevate the quality of artists' creative experiences and the artifacts they make. Yet most current CST evaluations overlook temporal and social aspects of tool use. To address this gap, we present a longitudinal, group-based CST evaluation through a three-week deployment of ArtKrit, a computational drawing tool that supports disciplined drawing. Nine digital artists, organized into three communities of practice, completed weekly "master studies" alongside a researcher-artist. Our results show users' evolving relationships with ArtKrit over time—from early experimentation to selective incorporation or misuse—alongside changes in their ways of artistic seeing. These changes unfolded within artist support networks that fostered confidence and creative safety, and validated individual expression. Overall, our findings suggest that CST evaluations can—and should—be designed as opportunities for meaningful artistic engagement rather than purely extractive measurement exercises.

CCS Concepts: • Human-centered computing → Field studies; • Applied computing → Fine arts.

ACM Reference Format:

Anonymous for Submission. 2026. Artistic Practice Opportunities in CST Evaluations: A Longitudinal Group Deployment of ArtKrit. In *Proceedings of 2026 ACM SIGCHI Conference on Designing Interactive Systems (DIS '26)*. ACM, New York, NY, USA, 24 pages. <https://doi.org/10.1145/nnnnnnn.nnnnnnn>

Author's Contact Information: Anonymous for Submission.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

© 2026 Copyright held by the owner/author(s). Publication rights licensed to ACM.

Manuscript submitted to ACM

Manuscript submitted to ACM

1 Introduction

Since establishing a Creativity Support Tools (CSTs) research agenda over 25 years ago [49], human-computer interaction (HCI) researchers have grappled with how to most appropriately evaluate them. One challenge is that the goals of CSTs vary widely: as noted in a 2005 National Science Foundation (NSF) workshop organized by Shneiderman et al., “the problem of developing CSTs is one in which one must first decide in which intersection of the n-dimensional taxonomy one wishes to study and work” [27]. The most common directions, as shown in a 2019 survey of CSTs [21], have been developing novel research prototypes for idea generation and artifact production. As such, researchers have used surveys such as the NASA-TLX (Task Load Index) [25] or the Creativity Support Index [12] to operationalize their definitions of “creativity”—often artifact- and process- centric—and justify that their tools help users achieve their “creative” goals.

Yet, the authors of the 2005 CST workshop report that “the first objective is to enhance the *personal experience* of the person who wants to be creative,” (emphasis added) while improving “outcomes and artifacts” and “process” are second and third objectives, respectively [27]. Most recently, Cox et al. highlight this oversight: beyond measures of productivity, HCI researchers should evaluate wholistic, *user-centric* aspects, such as how CSTs promote self-reflection and emotional well-being [46].

The shift to evaluating how CSTs may support the experiences of *artists*, and not just the creation of *artifacts*, shares goals with Research through Design methodology [23], which generates knowledge through reflective engagement with design artifacts, particularly towards speculative, diverse, and sometimes convoluted encounters [16]. Similarly, researchers deploy technology probes not to measure tool effectiveness in controlled lab settings, but rather to “find out about the unknown” in real-world contexts [28]. Commenting on the long-standing challenge of evaluating CSTs, Remy et al. call for researchers to consider longitudinal, in-situ studies, coming together as a community to develop a toolbox of methods [45].

This paper presents one such method for CST evaluation. Shaped by a belief that CST evaluations should surface generative insights about the creative practices and proclivities of artists who use them, we conducted a three week longitudinal evaluation of ArtKrit [41] with nine experienced digital artists. ArtKrit is a Krita [20] plug-in that supports disciplined drawing, the process of studying and replicating reference images, also called “master studies.” ArtKrit breaks down reference images into composition, value, and color scaffolds, offering computational guidance and feedback at each step¹. Our goals were not to demonstrate ArtKrit’s usability through statistical reproduction or quantify the improvements in reproducing master studies gained through using ArtKrit over a period of time. Instead, we focused on a qualitative observation (supplemented with quantitative data logging) of how computational scaffolds may alter digital artists’ workflows, relationships, and ways of artistic seeing and interpretation over time.

Motivated by Chung et al.’s theory of artist support networks [13], and given the prevalent critique of the lack of group-based CST evaluations despite creativity being a social activity [3], we grouped our nine participants into three communities of practice [18] on a shared Discord server. During each week of the three-week study, each community of practice collectively voted on a master study, which all members then drew individually. A member of the research team who was also an experienced digital artist joined each community of practice, drawing and posting on Discord. Our longitudinal, group-based study addresses a gap in prior CST evaluations that have largely relied on short-term, artifact-centric studies where participants do not know or interact with each other.

Our study aimed to answer three research questions:

¹For more details, please see the UIST ’25 publication [41].

- 105 (1) How do practitioners use ArtKrit in the wild?
- 106 (2) How do practitioners’ relationships with new creative software change over time?
- 107 (3) How do artist support networks shape practitioners’ experiences and processes?
- 108

109 We conducted a thematic analysis on collected weekly diary entries and semi-structured exit interviews, situating
110 our results in existing CST theory, paying special attention to emotional well-being and artist-centric benefits [46]. We
111 found that participants often deferred to ArtKrit’s suggestions for matters of composition and value, but remained firm
112 in their own color choices even when they did not match the reference image. A quantitative analysis of usage logs
113 revealed that engagement with ArtKrit decreased over time: participants moved from exploratory epistemic actions
114 to confidently choosing features that matched with their existing workflows. Creating artist support networks in our
115 evaluation resulted in feelings of creative safety; participants celebrated and were inspired by not how accurately their
116 peers recreated the master studies, but rather their unique personal adaptations.

119 Finally, we reflect on methodological implications from this longitudinal, group-based evaluation, where researchers
120 also participated as artists. We argue that CST evaluations can—and should—be structured as opportunities to commission
121 participants to meaningfully engage in their artistic practices. Specifically, by explicitly and intentionally setting a
122 study’s normative ground, researchers can co-create a space for participants to feel artistically and emotionally inspired
123 and supported, a step towards negotiating and lessening the power differential between researchers and “human
124 subjects.” Rather than “extracting knowledge” from artists, we hope our study can serve as one example in a broad
125 “toolbox” [45] of how CST evaluations can be generative for not just researchers, but also participants—participants
126 who may otherwise be frustrated that their unique, lived experiences extend beyond a narrow frame of knowledge
127 prescribed by research norms [6].
128
129
130

131 2 Methodology

133 To answer our research questions, we conducted a longitudinal, mixed-methods user study. Nine participants were
134 grouped into three communities of practice, each including a member of the research team, and used ArtKrit to replicate
135 one master study each week. Participants corresponded with each other and the research team on Discord. In total, the
136 study resulted in 36 artifacts and over 50 hours of ArtKrit usage data.
137
138

139 2.1 ArtKrit Improvements

141 To support real-world use during our longitudinal study, we implemented targeted system improvements for ArtKrit.
142 First, we reduced latency for adaptive composition line generation, especially on lower-memory machines, by tran-
143 sitioning the adaptive grid feature (the only feature that uses AI models) from a locally hosted model to an online
144 pipeline using Replicate. This change allowed participants to use ArtKrit’s adaptive lines without requiring specialized
145 hardware for local runs, which previously required sufficient GPU support and higher memory availability that was
146 not feasible for all participants’ devices. Second, we integrated incremental Krita saves and usage logging, recording
147 interaction events such as when participants requested feedback for composition, value, or color, along with the full
148 Krita document files saved at corresponding timestamps. These logs provided quantitative traces of tool use over time
149 that we utilized to analyze our participants’ drawing processes.
150

152 We also made minor quality-of-life improvements to ArtKrit. Toward more control, users can now regenerate
153 composition lines by manually moving, adding, or subtracting detected key points, without needing to first prompt
154 DINO and SAM. The composition tab additionally enables users to pop out and zoom their reference image. Finally, in
155
156

the color section of the tool, users can choose the background color (previously fixed as white) of the color separation feature, which also can be popped out and zoomed.

2.2 Participant recruitment

Researcher	Participant	Years of Experience	Age	Digital Tools Used	Artist Identity (Self-Described)
Researcher 1	1A	15	27	Photoshop, Procreate	Artist
	1B	15	31	Autodesk Sketchbook, Pixelorama	Hobbyist
	1C	10	22	Procreate	Hobbyist
Researcher 2	2A	8	20	Clip Studio Paint, Photoshop	Professional
	2B	15	22	Clip Studio Paint, MS Paint, Photoshop	Hobbyist
	2C	6	19	Procreate	Hobbyist/Professional
	2D	6	18	Krita, Procreate	Professional
Researcher 3	3A	7	19	Clip Studio Paint	Hobbyist/Aspiring Professional
	3B	4	20	Procreate	Hobbyist

Table 1. Participant demographics and self-described artist identities.

We recruited nine artists (five women, one man, three non-binary, and one who preferred not to disclose gender) through personal networks, with each researcher largely recruiting participants in their own group. Most of our participants came from snowball sampling where a participant would ask their friend(s) to join the study. All participants were experienced digital artists, though varied in their choice of tools and self-perceived artist identity. Table 1 shows participant demographics.

We split participants into small communities of practice to maintain a shared sense of cultural identity and context [31]. Group one consisted of three CST researchers who were also digital artists; Researcher 1 also shared this identity. Group two consisted of students at an art college—two were roommates, and one of the roommates was friends with the other two—that Researcher 2 had no prior relationship with, but recruited through mutual friends. Group three consisted of an art major at a non-arts focused college and their friend; Researcher 3 was their peer. Group three originally also included an art-major participant who was a friend of Researcher 3’s, but withdrew before completing drawings.

Following Jacobs et al. [29], we compensated participants at rates comparable to commissions; specifically, each participant was paid \$250 USD for three artworks.

2.3 Procedure

Our study design is informed by Long et al. [38], who found that user perceptions of AI tools suffer from novelty effects and undergo a familiarization phase. Figure 2 shows our evaluation timeline; we conducted the IRB-approved study November to December 2025. While we planned for participants to complete three master studies using ArtKrit over



Fig. 2. Overview of our evaluation timeline.

the course of three weeks, in practice the study ran 4-5 weeks due to onboarding and some participants completing Week 3 late as it ran into winter holidays.

First, during virtual onboarding, each researcher helped their group members install ArtKrit on their own computers. The researcher then conducted a 45-minute pre-study interview exploring participants' current practices around completing master studies, sharing art, and giving feedback.

During each week of the study, each group voted on which reference image to draw, voluntarily shared works-in-progress or other reflections on Discord, reacted to each other's images, and then spent approximately ten minutes completing a private diary entry once their image was done. The diary entry asked about participants' open-ended experiences with ArtKrit and their Discord group. Each facilitating researcher also actively participated by sharing their own artworks and conversing with their group throughout the study, though researchers did not complete diary studies, interviews, or submit usage logs.

Finally, at the end of Week 3, each researcher conducted individual exit interviews with their group members. To see if participants' digital drawing practices, relationships with ArtKrit, or ways of interpreting master studies changed, the exit interviews repeated many questions from the pre-interview, and also included participant-specific elaborations of their behaviors over the past three weeks. Exit interviews also lasted approximately 45 minutes.

2.4 Data Analysis

Data collected and analyzed include usage logs, weekly diary entries, Discord messages, and pre- and post-interviews. To qualitatively analyze the interview transcripts and diary entries, we conducted a deductive thematic analysis [9]. We chose a deductive, as opposed to inductive, thematic analysis to explore our pre-existing research questions and to find evidence supporting existing CST theory, such as artist support networks [13], dialectical activities [53], and a lens of power [37]. The first author coded all data given these themes, and through open discussions and weekly meetings, the research team reached a consensus.

For quantitative data, we analyzed system usage logs collected during ArtKrit's three-week deployment, including JSON interaction logs and their corresponding system-logged canvas images. We first cleaned user-uploaded data by removing records outside the three-week study period, and then grouped the data per study group. Next, we selected all user-initiated actions and categorized them as composition-related, value-related, or color-related. Finally, we conducted exploratory data analysis in Python to examine trends and usage patterns, including total action counts, the frequency and percentage of each action category, and changes in action usage across weeks and study groups. We additionally analyzed the appearance and disappearance of unique actions over time to understand how participants' interactions temporally evolved.

2.5 Positionality

The first, third, and last authors participated in the study. We take inspiration from auto-biographical methods in HCI [15] that encourage reflexivity. Originally, we decided to participate mainly to try to reduce attrition through establishing greater rapport with participants, but we found that our experiences in the study also changed our relationships to ArtKrit. While we acknowledge that our actions and reactions cannot be entirely disentangled from those of participants, to retain a narrative that highlights the contributions of participants, we do not intermix our data or impressions with theirs. Instead, we offer first-person accounts and reflections of our experiences below.

First author (Researcher 3): I used Procreate to draw and I uploaded my works in progress to Krita and ArtKrit whenever I was looking for feedback or support. Like many of the other participants, I found the rule of thirds grid the most helpful and used it frequently to create the initial structure of my piece. Previously, I believed that I would've preferred to use adaptive composition lines and composition feedback compared to pre-set grids, so I was surprised by how helpful rule of thirds was for me in terms of correct composition. Before using ArtKrit in our evaluation, I would've said that composition, value, and color were equally important. After doing this study I think that composition and value are more significant to the piece and color can be changed according to artist preference. I found myself choosing colors that I liked rather than striving for color similarity with the reference. I also found myself misusing the tool by using the color separation feature for value blocking to show me where lighter and darker areas were. Finally, I would echo something that 1C mentioned, which is that ArtKrit slowed down the initial steps in my workflow, to make future steps easier and less overwhelming.

Third author (Researcher 2): In the beginning of the study, I did all of my work in Krita. However, towards the end of the first week, I switched to drawing in Procreate and uploading my works in progress onto Krita, similar to my fellow authors. This approach streamlined my drawing process while also getting the necessary support from ArtKrit. I found the composition functions to be the most helpful. I would first generate an adaptive grid and then move the lines around into a position I found the most helpful. This process gave me a good starting point for scaffolding, while also giving me control to create a grid best suited for my needs. I usually don't focus on values when drawing, but all of my group members emphasized its importance for any piece. Thus, I paid more attention to this aspect and realized that my values tend to be inaccurate. This pushed me to use the value functions more, such as the different filters and kernels. Through these interactions, I discovered a new way to look at drawings and a novel way to use ArtKrit in order to improve my artistic skills. I felt similarly to 2C and 2A who emphasized how the study and its social aspects inspired them to explore different stylistic techniques and ways of using ArtKrit.

Last author (Researcher 1): Like the first author, I stayed true to my workflow, drawing in Procreate/acrylic and horizontally moving to/from ArtKrit for scaffolds and feedback. I interpreted the master study of Week 1 as fan art of Jacinthe from Pokémon: Legends ZA as I often do this in master studies to still feel like I'm retaining creative agency; I didn't consciously know at the time it would set a norm! I was inspired by my group's reflections of their creative misuse to push ArtKrit to its limits to see how it would handle traditional art for Week 2. It was tricky to control for ambient lighting while photographing the painting; at first, ArtKrit told me my painting was too red (I had a red underpainting), but in retrospect I think I overcorrected due to the warm lighting of the photo. Even though I've used ArtKrit for a long time even before the study, I found my mental framing of it change as a result of my participation. Like what 1C said in their exit interview, ArtKrit was most useful in showing *spatial relationships* of values and color. I realized I now look at and evaluate drawings wholistically, with all aspects of composition, value, and color combined, even if I focus on separate parts during drawing.

As the senior author, and as a researcher in a group with participants who were all more junior HCI researchers, I was always conscious of my position of power and tried not to overstep any bounds (in the exit interviews, group members reported they didn't feel uneven power dynamics). Mainly, though, I was happy to have an excuse to do my own creative practice, and to be supported by artists who felt more like a community of friends than research participants.

3 Findings

Below, we report the results of our three research questions; we supplement our thematic analysis narrative with quantitative usage log analysis. Each subsection's findings are situated in existing CST theory and ends in design implications for other CST researchers and developers.

3.1 RQ1: How did participants use ArtKrit in the wild?

On average, participants spent 2 to 3 hours, and up to 6 hours, completing their weekly master studies, though their engagement with ArtKrit was less than drawing time (Table 2). Typically, participant workflows would start with using grid lines to get a rough sketch or line art of the composition of the reference image. Then, with the use and misuse of various features in color and value, including value and color comparison, color separation, and lasso fill, participants would continue to either value blocking or color blocking. Following steps consisted of working with value and color in no particular order to look at spatial groupings, "vibe check" their piece, and make continuous adjustments until the piece felt complete. For example, Figure 3 shows Participant 1A's three-hour drawing session of using ArtKrit in Week 1. They began by setting up simple grids and gradually sketched and rendered the entire piece. The participant then painted over the grid, integrating it into the image while adjusting head positions using the rule of thirds and refining value rendering. Towards the end, they made finer adjustments, such as shrinking the head, and completed the rendering after incorporating value feedback. Figure 4 shows final results from Group 1, Figure 5 shows Group 2, and Figure 6 Group 3.



Fig. 3. Participant 1A's three-hour Week 1 drawing session in ArtKrit, showing progression from initial grid setup to final rendering. Key actions include sketching, grid integration, updating previews, refining value, and adjusting proportions.

3.1.1 Productive Friction. Kreminski et al., in *Reflective Creators*, call for CSTs intentionally designed to elicit reflection and to support "thoughtful creative practices and practitioners" [34]. A key mechanism through which reflective creators prompt reflection is design friction, which interrupts otherwise automatic creative action and invites users to

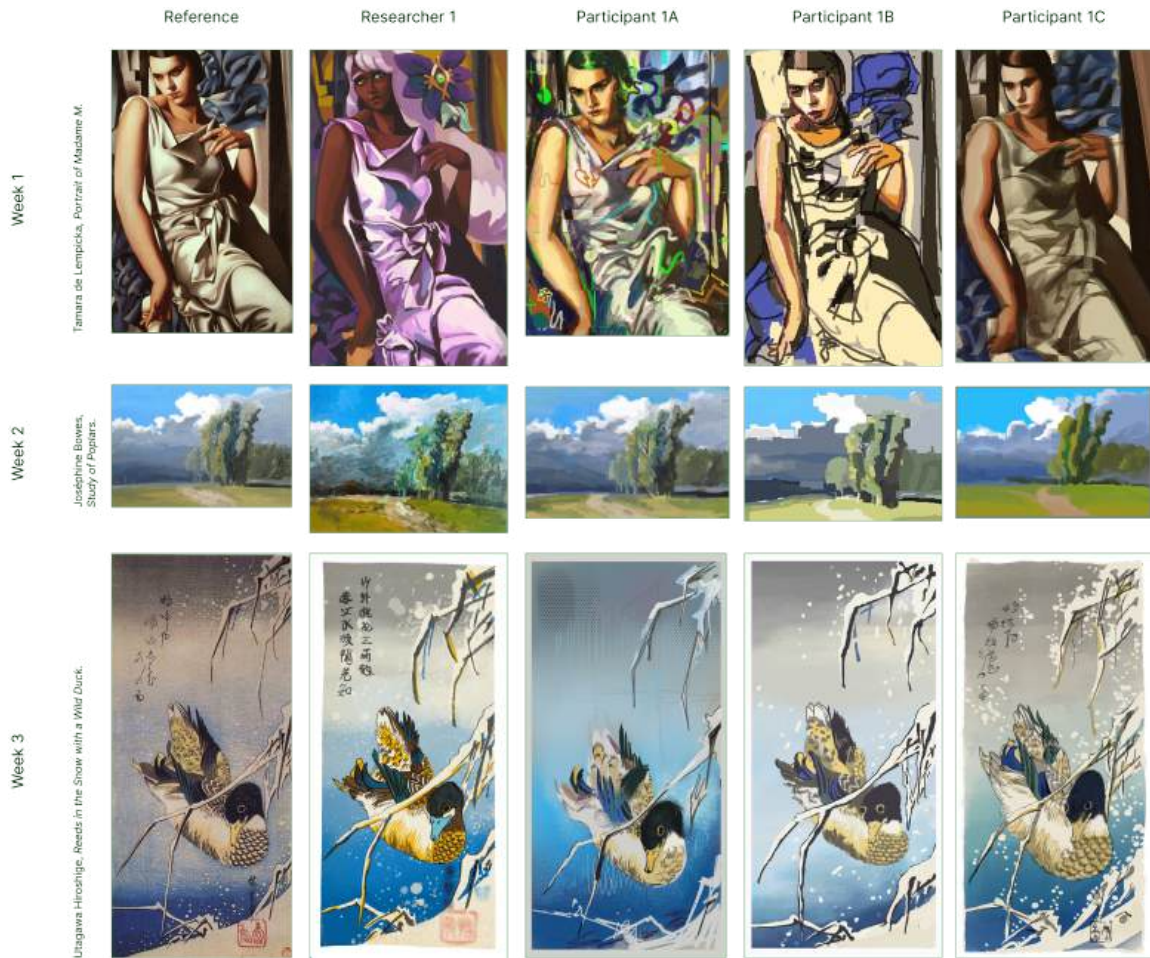


Fig. 4. Drawings produced by the participants of Group 1. Week 1 reference: *Portrait of Madame M.*, Tamara de Lempicka. Week 2 reference: *Study of Poplars*, Josephine Bowers. Week 3 reference: *Reeds in the Snow with a Wild Duck*, Utagawa Hiroshige.

reconsider their choices. We observed multiple instances of reflection emerging from friction introduced into artistic actions, decisions, and interactions with the tool.

One way friction was evoked was through the choices the tool left open to users, rather than the tool making explicit changes. As 1A explained, using ArtKrit required active self-assessment: “you have to actively reflect and be like, am I being supported? Because it’s kind of like a thing that’s happening [on the sidebar] as opposed to an explicit mark that has changed because you’re using ArtKrit.” This ambiguity prompted users to reflect on when to follow the tool’s suggestions and when to assert their own agency.

We observed cases in which participants chose to override ArtKrit’s feedback in favor of their own preferences or creative intentions. For example, after receiving color feedback, both 1C and 3B decided not to adopt the tool’s suggestions. 1C noted, “When I was using ArtKrit and getting feedback on the colors, I wasn’t usually taking it, because I know what colors I like, even though it doesn’t match the reference image.” Similarly, 3B described ignoring a prompt

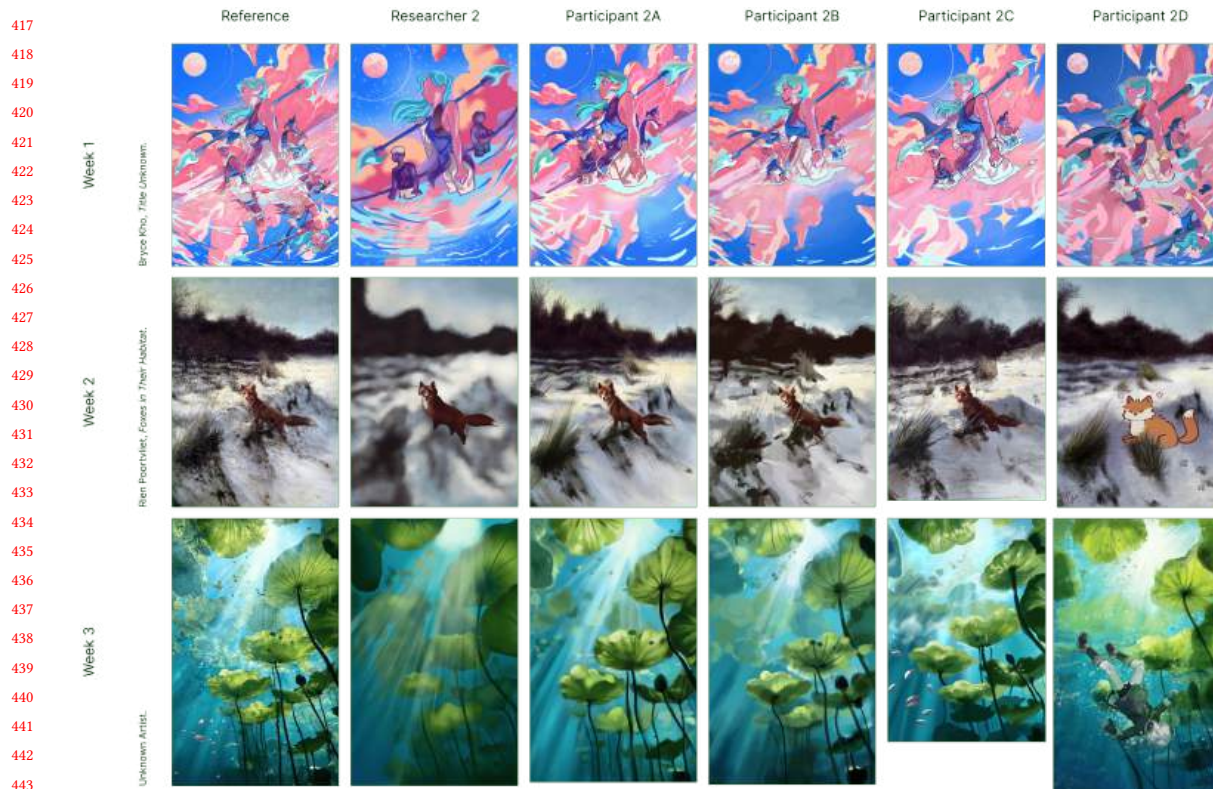


Fig. 5. Drawings produced by the participants of Group 2. *Week 1 reference: Title Unknown, Bryce Kho. Week 2 reference: Foxes in Their Habitat, Rien Poortvliet. Week 3 reference, Title Unknown, Artist Unknown.*

because replicating the figure exactly wasn't their intention for Week 3: "the feedback it gave me was like, you don't have enough yellow. And I was like, yeah, I know that. Because like my finger isn't wearing a yellow shirt, so obviously." In these moments of friction, artists made a reflective decision about their judgment taking priority over the tool's judgment. This tension is especially salient in the context of master studies, where the goal is often to replicate an existing piece. However, we observed artists acknowledging and ignoring ArtKrit's feedback when replication did not contribute to their learning or artistic goals, such as when they were instead focused on personal expression.

At the same time, we also saw cases where participants prioritized ArtKrit's feedback. 3B reflected, "I was a little bit more dismissive of some of the things that ArtKrit was telling me that I was doing wrong, but I would say that values and composition were definitely still in effect." When participants acted on this feedback, they reflected on whether doing so supported their artistic learning. 1B, for example, adjusted the placement of their duck after receiving composition feedback, but noted that "it didn't require [them] to actually understand visually where the duck was; [they] just took what the tool was telling [them] at face value," which made the experience "feel like tracing." Across these examples, we see a recurring tension between questions of support ("am I being supported by the tool?") and growth ("am I actually improving as an artist?").

We observed a pattern across these instances: we saw user agency take priority in color, and tool suggestions take priority in value and composition. Participants such as 3A, 3B, and others emphasized maintaining the core essence,



506 Fig. 6. Drawings produced by the participants of Group 3. Week 1 reference: *Butterflies*, Fujishima Takeji. Week 2 reference: *View from*
507 *Massaki of Suijin Shrine*, Utagawa Hiroshige. Week 3 reference: *illustration in Summer Classroom*, Kenji Tsuruta.

508
509 or most significant elements, of the reference image as their goal when completing master studies. This leads us to
510 ask: were participants more willing to heed ArtKrit’s feedback on composition or value because they believed these
511 elements were most essential to maintaining the essence of the original?
512

513
514 **3.1.2 Dialectical Activities.** Zhang defines dialectical activities as activities whose values are rooted in the intrinsic
515 experience of engaging in the activity itself, rather than in external outcomes [53], citing examples such as art-making
516 and other creative pursuits. Zhang further calls for ecosystems that simultaneously produce desired goods and services
517 while also promoting people’s sustained engagement in dialectical activities. In our study, we observe instances in which
518 ArtKrit appears to shape and intensify dialectical engagement within the art-making process. Rather than supporting
519
520 Manuscript submitted to ACM

instrumental goals (e.g., producing an accurate image), ArtKrit introduced moments of reflection, emotional response, and perceptual consideration that became meaningful as part of the act of creation itself.

ArtKrit’s features encouraged practices rooted in learning visual fundamentals, which improved participants’ ability to artistically see. One of ArtKrit’s most frequently used features across all groups was toggling the rule-of-thirds grid. Several participants noted that simply pressing a button to see grid lines, rather than drawing them manually, prompted them to engage in practices they might otherwise skip or overlook. Similar “single button press” actions included viewing a black-and-white version of the canvas (for values) instead of manually applying a grayscale filter, and hovering over regions in the color separation tool rather than manually color-blocking. 2A explained, “I don’t usually compare values. It’s kind of difficult when you’re drawing from real life, or it’s just kind of tedious, but ArtKrit made it pretty easy to see all these different aspects of your reference that you wouldn’t normally think about.” 2C added, “I wouldn’t really think much about it, but looking at it now, I’m kind of noticing the style of the artist and the colors, how the colors relate to each other, and how they really put time into the composition of the piece... before, I would just mindlessly draw without thinking about any of those things.” 1A described this process as “eating your vegetables”: “Having [ArtKrit] do [things] for me in a way that fed me my vegetables, I see why this is a nourishing thing to do.” ArtKrit’s support in improving seeing highlights one type of dialectical value that is intrinsic to artistic creation.

Similarly, ArtKrit can provide emotional and cognitive support in the art creation process by making complex images feel more approachable and less overwhelming. Alleviating these barriers is another way that using ArtKrit can support dialectical practices. 3A reflected on Week 1’s image, explaining that ArtKrit helped them “understand the fundamentals a lot better, rather than being overwhelmed” by the many butterflies in the image. 1A described a similar experience for Week 3’s duck study, highlighting how the intersection of the adaptive grid lines encompassed a specific region of feathers that they could focus on rendering, rather than feeling overwhelmed by the whole duck. 2A emphasized that ArtKrit gave them the tools to “dissect” color and value, enabling them to engage more mindfully with individual elements of the piece.

3.1.3 Surfaced Limitations. Through our longitudinal deployment of ArtKrit, we also identified several limitations. Because ArtKrit is designed for replicating reference images, it is less effective for original work or intentional deviations from a reference. For example, 3B had to dismiss some of ArtKrit’s suggestions due to intentional differences from the reference in Week 3. As a workaround, Researcher 1, who also added original aspects to their studies, edited the reference image to reflect desired changes. Several participants also noted performance issues dissuading them from using features, particularly the slow generation of adaptive composition grid lines, which took between thirty seconds and one minute on average on the Replicate servers. Additionally, 2D did not use the color feature because they found it complicated and felt that the value feature was sufficient. Participants offered suggestions for improvement as well: 1B expressed interest in a more localized value/color selection feature, which would allow comparisons of smaller regions rather than the entire canvas. As part of our iterative design process and commitment to improving the tool with users, we have updated ArtKrit to add this functionality.

3.1.4 Design implication: Integrated simplicity bests novel algorithms. Throughout the study, we observed that sometimes, the most technically simple features (i.e., rule of thirds grid) had the greatest influence on artistic practices, especially when compared to novel AI-assisted algorithms (i.e., adaptive grid lines). Even though all participants were capable of manually drawing a rule of thirds grid in about a minute, the ability to press a button to instantly generate one was enough of a savings in manual labor that participants could instead focus on observation and interpretation, rather than

573 spending time creating scaffolds. Just because a feature is known, easy to implement, and not a technical advancement
 574 does not mean it will not be impactful to creative practices.
 575

576 3.2 RQ2: How did relationships with ArtKrit change over time? 577

578 As participants used ArtKrit over the course of three weeks, their relationship with the tool evolved. This shift is
 579 visible in Table 2 and Figure 7, where participants initially engaged in more exploratory use and, by the end of the
 580 study, settled into using the features that best supported their individual workflows. During this exploratory phase,
 581 participants experimented with different parts of ArtKrit to understand how it could support their existing practices,
 582 foregrounding epistemic over pragmatic action [51]. Early on, 1A described “going around the different tabs” and
 583 observing the cognitive impact of various features.
 584
 585

	Avg. Active Hours (per user)				Avg. Active Days (per user)			
	3-Weeks Avg.	Week 1	Week 2	Week 3	3-Week Avg.	Week 1	Week 2	Week 3
1A	1.64	2.91	1.78	0.22	2.91	2.00	2.00	1.00
1B	1.97	2.07	1.65	2.21	2.33	2.00	2.00	3.00
1C	0.70	n/a	0.55	0.86	2.00	n/a	2.00	2.00
Group 1 Avg #	1.44	2.49	1.33	1.10	2.00	2.00	2.00	2.00
2A	1.73	1.43	2.73	1.02	1.67	2.00	2.00	1.00
2B	1.72	0.66	2.87	1.62	2.00	2.00	2.00	2.00
2C	0.38	0.42	0.04	0.68	1.33	1.00	1.00	2.00
2D	3.73	6.10	4.17	0.93	2.00	3.00	1.00	2.00
Group 2 Avg #	1.89	2.15	2.45	1.06	1.75	2.00	1.50	1.75
3A	3.87	5.15	3.30	3.16	2.33	3.00	2.00	2.00
3B	1.23	1.75	1.87	0.07	2.00	2.00	2.00	2.00
Group 3 Avg #	2.55	3.45	2.59	1.61	2.17	2.50	2.00	2.00
All Participants	1.89	2.56	2.11	1.20	1.94	2.14	1.79	1.90

600 Table 2. Average weekly active hours and days per participant in ArtKrit. Note: Week 1 data logs for Participant 1C are unavailable
 601 because they were not uploaded and were accidentally deleted by the participant.
 602
 603

604 **Quantitative Evidence: Decline in Active Hours Spent Over Weeks**

605
 606
 607 To understand users’ active usage time with ArtKrit over three weeks, we computed active hours and days for
 608 each user and group from system logs. We defined active time as consecutive actions within a 30-minute window
 609 (e.g., users might be working on drawing between system interactions) and an active day as any calendar day on
 610 which a participant performed at least one action. Table 2 shows the elapsed active hours and days each participant
 611 spent working with ArtKrit over the three weeks. Across all nine participants, they spent an average of 1.89 active
 612 hours over 1.94 calendar days per week using ArtKrit.
 613
 614

615
 616 Across weeks, the average time spent with ArtKrit shows a clear declining trend in two groups: Group 1 (from
 617 2.49 hours to 1.33 hours, then to 1.10 hours) and Group 3 (from 3.45 hours to 2.59 hours, then to 1.61 hours).
 618 The decline in usage time of ArtKrit can potentially be attributed to exploratory actions in using a new tool:
 619 participants engaged more deeply with ArtKrit in the first week to understand how the tool could fit into their
 620 existing workflows, after developing mental models of how it could be used, they utilized only features they knew
 621 to be helpful for them [51]. This idea is expanded upon in 3.2.1.
 622
 623
 624

Quantitative Evidence: Decline in Total Actions Over Weeks

Beyond decreased active usage time, participants' interactions with ArtKrit also declined over the course of the study in terms of both the total number of actions taken and the breadth of unique actions used. As shown in the top chart in Figure 7, the average total number of actions dropped from 76.9 in Week 1 to 36.1 in Week 2 and further to 30.8 in Week 3, representing a 53.1% decrease from Week 1 to Week 2 and an overall 59.9% decrease from Week 1 to Week 3. Similar to trends observed in earlier sections, this reduction likely reflects a shift away from exploratory behavior toward more targeted use of ArtKrit features. This overall decline was consistent across groups, as shown in all three left charts in Figure 8, though the trajectories differed. Groups 2 and 3 exhibited steady, monotonic decreases across weeks. Group 2's average total actions fell from 46.8 in Week 1 to 22.2 in Week 2 (a 52.6% decrease) and further to 19.5 in Week 3, amounting to a 58.3% decrease relative to Week 1. Similarly, Group 3's actions declined from 117.5 in Week 1 to 59.5 in Week 2 (a 49.4% decrease), followed by a sharper drop to 26.5 in Week 3, corresponding to a 77.4% overall decrease. Group 1 followed a slightly different pattern, with a sharp reduction from 96.5 actions in Week 1 to 39.0 in Week 2 (a 59.6% decrease), followed by a modest rebound to 48.7 actions in Week 3; nevertheless, this still represents an overall decrease of 49.5% from Week 1 to Week 3.

In addition to taking fewer actions overall, participants also engaged with a smaller set of unique action types over time. After the first week, users performed fewer distinct actions on average when using ArtKrit. Across all groups, participants used an average of 21.67 unique actions in Week 1 (23 for Group 1, 19 for Group 2, and 23 for Group 3). In Weeks 2 and 3, this number decreased to an average of 17.67 unique actions, reflecting a reduction of approximately four action types (a 19% decrease). Taken together, these patterns suggest that participants gradually developed more streamlined and personalized workflows with ArtKrit, relying on a smaller but stable subset of features as they gained experience across tasks.

Over time, participants developed an understanding of which features supported specific aspects of their workflow, and learned to draw on different parts of ArtKrit (composition, value, and color) as needed, such that "when they cognitively needed that support, they knew which tab of ArtKrit to go to." (1A). Similarly, mental models of ArtKrit's potentials and limitations emerged as the study progressed, suggesting sustained and effective use across its feedback dimensions rather than increasing reliance on a single aspect. For example, when using adaptive composition lines, participant 1B initially thought that text inputting a single object would cause lines to appear only through that object, but later realized that prompting two objects was necessary to generate lines between them. Over time, participants' engagement with composition-, value-, and color-related features tended to converge toward a more even split, with each aspect receiving a similar amount of attention in their workflows.

Quantitative Evidence: Convergence Toward a Balanced Distribution of Action Types

Across the three action categories—color-related, value-related, and composition-related—we observed an initial imbalance in usage that gradually converged over time. As shown in the bottom chart in Figure 7, all participants' interactions were dominated by composition-related actions (48%) in Week 1, with color-related (27%) and value-related actions (26%) used substantially less. Over the following weeks, however, these differences narrowed. Color-related actions increased steadily, rising to 37% in Week 2 and 38% in Week 3, while composition-related actions declined to 44% in Week 2 and 35% in Week 3. Value-related actions remained relatively stable throughout,



Fig. 7. Average number and percentage of user actions by category (color, value, composition) over three weeks.

hovering around 26% across all three weeks. By Week 3, the distribution of actions converged toward a more even split, indicating a process of gradual equalization in how users appropriated different feedback dimensions over time, rather than continued reliance on a single dominant category.

Similar gradual equalization patterns are observed across study groups, as shown in the three right charts in Figure 8. In Group 1, early usage was strongly imbalanced across action categories (18% color, 22% value, and 60% composition in Week 1). Over time, both color- and value-related actions increased (color from 18% to 31% to 43%; value from 22% to 23% to 24%), while composition-related actions decreased substantially from 60% to 46% and then to 32%, resulting in a more balanced distribution by Week 3. A similar convergence pattern appears in Group 2, where color-related actions—the least represented category initially—rose from 11% in Week 1 to 30% in Week 2

729 and 37% in Week 3, while composition-related actions declined from 58% to 41%. Group 3 followed a different but
730 still equalizing trajectory: color-related actions decreased from 46% to 26% over time, while value-related actions
731 increased from 25% to 40%, and composition-related actions rose modestly from 30% to 34%.
732

733 Taken together, these trends suggest that ArtKrit’s three-aspect design—composition, value, and color—supports
734 a workflow in which users gradually engage more evenly across aspects over time. Initial imbalances may stem
735 from participants’ prior drawing habits, stylistic preferences, early exploration of ArtKrit’s novel composition
736 algorithm, or the specific reference images used in a given week. As participants continued using ArtKrit, they
737 increasingly converge toward more balanced usage across three aspects rather than reliance on a single dominant
738 category.
739
740
741

742 As participants worked with a range of paintings and artistic styles throughout the evaluation, we observed that
743 ArtKrit was used differently depending on the complexity of the reference. For more complex paintings, such as the
744 Group 3’s Week 1 *Butterflies* study, ArtKrit was integrated directly into the process of placing elements onto the
745 canvas. In contrast, for simpler pieces, such as the Group 1’s Week 2 *Study of Poplars*, ArtKrit functioned more as a
746 post-hoc “sanity check.” For example, 1A used ArtKrit to test their color accuracy only after that portion of the piece
747 was already complete, framing it more as a game to check their color matching abilities. Participants exercised agency
748 in deciding both the timing and extent of feedback, engaging with ArtKrit at different stages of their workflow. Overall,
749 simpler pieces led artists to prioritize the cohesion of the whole composition, whereas more complex pieces encouraged
750 attention to individual sections. This was evident in Group 2’s complex Week 1 image, which 2B described as a “paint
751 by numbers” that they were mindlessly copying because the colors were so complex.
752

753 We also observed instances of what might be characterized as creative misuse, where participants employed ArtKrit
754 in ways beyond its original design intent [36]. One such example involved 1A using ArtKrit as an intuition training
755 tool for checking their color matching skills. In another instance, while working on Week 1’s *Portrait of Madame M.*,
756 1A incorporated ArtKrit’s neon green grid lines directly into the artwork as their goal was to “respect the medium
757 of ArtKrit,” describing them “as if there were green laser lights in the room with her.” 1C stated that ArtKrit’s most
758 helpful feature was the neon contours of similarly grouped regions generated by the value and color feedback panels.
759 But instead of using the contours overlaid on the reference image as comparative feedback against their canvas, 1C
760 repurposed these visually salient elements to guide their composition. They noted that the shapes of the neon blobs
761 helped them see sharper regions of color, and used it when color blocking their drawings.
762

763 Through participating in the evaluation, participants reported changes in their artistic practices, preferences, and
764 skills that they intended to carry forward beyond the study. For example, 2D discovered new brushes they wanted to
765 continue using: “[The master studies] made me use a lot more brushes, like they put me out of my comfort zone. I
766 actually found a bunch of new options [that] I’m probably going to be using a lot in the future.” Similarly, 1A described
767 feeling more prepared to pursue physical plein-air landscape paintings after completing Week 2’s *Study of Poplars*.
768

769
770
771
772
773
774 **3.2.1 Design implication: Reaffirming the familiarization phase.** Our results, both qualitative and quantitative, suggest
775 that participants’ use of tools evolves as their familiarity with them increases. Over time, they begin to integrate the
776 tool into their existing workflows and practices. For example, 3B, a participant in an earlier evaluation of ArtKrit, noted
777 that in the first study they experimented more consistently with ArtKrit’s features as the study took place in a lab
778 setting and using ArtKrit was what they were there to do. In contrast, in the current study, 3B was more comfortable
779



Fig. 8. Average number and percentage of user actions by category (color, value, composition) over three weeks by each study group.

using their usual setup and was more selective about the features they found most useful. We offer evidence in support of Long et al.'s theory that users undergo familiarization phases with novel tools [38]. We believe that as epistemic, Manuscript submitted to ACM

exploratory actions account for Week 1’s high usage counts, CST evaluations that argue for a tool’s usability should consider actions after a familiarization phase to avoid novelty effects.

3.3 RQ3: How did artist support networks shape participants’ experiences?

Using ArtKrit within an artist support network [13] contributed to participants’ emotional well-being through positive interactions and encouragement. Consistent with theories of social support and communities of practice, these interactions functioned as expressive and affirmational support, reinforcing participants’ sense of belonging and legitimacy within the group [52]. Feedback and discussion within groups were almost entirely positive, with participants commenting on the uniqueness of each master study and individual expression rather than accuracy or similarity to the original. Some representative comments reacting to finished pieces in the Discord included “I like that soft line brush you used for the clouds” (2A reacting to 2D’s Week 1 painting) and “This is adorable, I love the outfit” (2C reacting to 3B’s Week 3 painting), in addition to the messages in Figure 1. Completing studies within a group was also grounding for participants. As Wenger describes, communities of practice “provide homes for identities,” allowing participants to situate their struggles and progress within a shared practice. 2C shared that they enjoyed having a space to talk with others about the pieces rather than getting “in their own head,” when they were struggling and heard 2A mention similar difficulties, it made them feel that their struggles were seen. 3B added that being part of such a group felt almost equalizing: when viewing professional artists’ work, it can be difficult for them to imagine how the piece was made, but seeing peers successfully complete the same studies was both reassuring and motivating. In viewing participants’ encouragement as a reception activity in a creativity support ecosystem [1], we highlight how peer feedback and affirmation extend beyond evaluation, functioning instead as social infrastructure that supports emotional resilience and creative persistence.

Erickson and Kellogg highlight the importance of systems that allow users to “see” one another, make inferences about others’ activities, and imitate one another, arguing that such social translucence enables environments to “support the same sort of social innovation and diversity that can be observed in physically based cultures” [19]. Through our longitudinal group evaluation, participants described becoming more motivated, more willing to invest effort, and more likely to explore and integrate new possibilities into their own workflows through observing and responding to others’ practices. For instance, 1C went back to and spent an additional three hours on their duck drawing in Week 3, citing they did not want it to “look bad” compared to their peers’ drawings, which may explain the increased average actions and active days shown in Table 7 and Group 1’s Week 3 bar in Figure 8.

We observed how participants were influenced in both managing- and teaching-type tasks in their artist support network [13]. For instance, for managing-type tasks, participants noted that seeing the stage others were at during the week motivated them either to get started or to continue working. Participants continually set and reacted to standards of effort and completeness: 2B mentioned that seeing some people “turn in” minimal pieces was reassuring, as it affirmed that even if their own piece was not amazing, it was still acceptable. 1B noted that “seeing what details other people chose to spend time on rendering definitely influenced [their] choices on what to spend time on, or how long to keep going,” effectively setting a benchmark. They also mentioned learning through observing others in teaching-type tasks: “I definitely picked things up from seeing other people do it, and then I think it also, like, gave me an idea of what quality bar to shoot for.” 1A similarly indicated that they wanted to put in effort because they saw others doing so, “not in a pressuring way, but in a co-working sense.” 2C, who initially relied on eyedropping colors, copied less over time and instead tried to “learn” more by observing others’ work.

885 A group-based evaluation also helped expand what was possible within the study. Many participants pointed
886 Researcher 1’s modified Week 1 master study (Figure 1) as a realization of what could be done, which led to increased
887 personalization and content changes in participants’ work in subsequent weeks. Under a lens of power [37], researcher
888 participation helped set a normative ground by example, rather than researchers explicitly enforcing or suggesting
889 practices. Beyond researcher involvement, participants’ own practices also inspired the practices of others. For instance,
890 1A’s inclusion of green grid lines in Week 1 inspired 3B to do something similar in week three, and 1B’s use of a
891 trackpad encouraged 1A and 1C to do their Week 2 studies with a trackpad as well. As 1A put it, “you don’t really see
892 the possibilities until you see somebody kind of pushing the possibilities.”
893
894

895 Lastly, all participants reported that they had fun participating in the study: “I had so much fun! I wanna do more
896 studies! I think a Discord group where people make art together is a very good way to do a study!” (1A); “Working
897 with the tool was really fun. It felt exciting to be part of making something new” (3A); and “I really had fun with it.
898 Yeah, I really enjoyed the study” (2D). Beyond completing study procedures, this underscores the often overlooked
899 aspect of evaluations including user-centric value, whether through positive emotions, increased self-confidence, or
900 overall well-being [46]. Our findings reveal the potential of CST evaluations to contribute positively to participants’
901 lives through lived experiences, highlighting the value of considering affective and experiential outcomes as part of
902 understanding how creative systems are experienced by users.
903
904

905
906 *3.3.1 Design implications: A push towards greater group studies.* These findings highlight the value of social and
907 collaborative structures in supporting creative work. Being part of an artist support network not only encouraged
908 emotional well-being through positive feedback and shared experiences, but also helped establish implicit standards for
909 motivation, effort, and the possibilities of practice. Artists create art, in large part, to react to the norms around them
910 [36], and a group-based study presents an opportunity for any participant to enact a new norm. We believe group-based
911 studies may expand the space of what is artistically possible while providing support and emotional-well being for
912 creative practices, and encourage CST researchers to look into adapting group-based methodologies to evaluate their
913 own tools.
914
915

916 4 Discussion

917 Here, we reflect and speculate on our evaluation methodology and its implications for future HCI research. While our
918 findings revealed that researchers have the power to set a study’s normative ground, some may view this intentional
919 setting with concern as it could “bias” participants. We offer a response, and argue that evaluations—not just the research
920 phase of tool design—are still sites for creative practice. We believe CST researchers should use their power, authority,
921 and financial resources to enact ad-hoc artist support networks during evaluations.
922
923
924

925 4.1 Setting normative ground versus biasing participants

926 In section 3.1.2, we argued that our longitudinal group evaluation of ArtKrit made space for dialectical activities
927 from which participants derived some intrinsic benefit. Participants’ behaviors were in part guided by the normative
928 ground researchers set, such as Researcher 1 turning a study into Pokémon fan art to show personal interpretation is
929 accepted—and even welcomed.
930
931

932 Traditional HCI evaluations, such as controlled in-lab studies, may frame norm setting as a threat to validity or
933 neutrality, with the potential to bias participants. However, we argue that norm setting is essential to CST evaluations—
934 particularly in evaluating CSTs with artistic goals, where users value process and expressivity within wide walls. In a
935

creative practice, norms shape what kinds of learning and expression are possible. Artists create art, in part, to respond to norms [36]. Design probes are not value neutral [22], and, echoing calls of defining what the *design* goals of a CST should be [21, 27, 45], HCI researchers should also define their *evaluation* goals. By offering examples that sit outside of participants' pre-conceived notions of acceptable master studies for a research study, researchers "biased" participants to expand their notions of the space available for creative opportunity. We hope this resulted in more meaningful creative engagements, such as how participants then commented on and supported everyone's unique touches, which bolstered creative safety and led to more creative risk taking.

Thus, we argue for moving towards thinking of CST evaluations not as data gathering in service of validating a tool's design goals, but rather as **infrastructure design**. In designing the structure of an evaluation, researchers have the opportunity to guide participants emotionally, epistemically, and practically through meaningful creative engagements. For instance, even though 2A said they would not post any of their master studies on social media since none of them were in their usual fan art style, they still cited developing transferable skills through participating in the study. Though spiritually similar to participatory design methods [8], which emphasize co-designing the *tool*, none of our participants aided in the design of ArtKrit. Instead, the *study* was a co-created experience. While researchers initially planned the structure (e.g. small communities of practice that draw the same master study each week), participants helped recruit one another into the community of practice, selected the master studies, and established feedback norms. With just researcher guidance, participants might not have been sufficiently motivated and influenced in their artistic choices, such as how the rest of Group 1 mirrored not the researcher's traditional painting, but 1B's track pad drawing during week two. Rather than a shared ownership of tools with participants, designing CST evaluations as infrastructure for artist support networks results in a shared ownership of a creative experience.

To frame this discussion under a lens of power [37], we distill two design implications:

- Creating CST evaluations as infrastructure for artist support networks and opportunities for meaningful creative engagement is a way researchers can mitigate the power imbalances they have with participants.
- By also participating as artists, researchers may intentionally set normative ground through demonstrating their own artifacts.

4.2 Generalizing to other CSTs

While the goals and affordances of ArtKrit dictated some of our study structure, we believe our methodological lessons learned generalize to other CSTs. A CST evaluation is one of the rare opportunities researchers have to evaluate real artists using their tools—under the Creativity Support Ecosystem [1] perspective, group-based longitudinal studies mean that researchers have some control over distribution and reception activities, not just creation ones. And the distribution and reception activities that occur in an evaluation are also parts of the creative practice: **a CST evaluation is not external to the design process; it is one of the sites where creativity is shaped, practiced, and made legible.** The participants researchers recruit, the tasks researchers delegate—all are ways to shape the resulting ecosystem and artist support network. It is our hope that CST evaluations consider this wholistic view, as the HCI community works towards creating generative sites of understanding human creative activity.

Granted, a longitudinal group-based evaluation is just one tool in the toolbox of potential CST evaluations [45], and evaluations should be picked to match the goals of both researchers and participants. A longitudinal group-based technology probe that aims to uncover emergent patterns of creative use may not be the best fit for proving metrics like increased efficiency or reduced cognitive load. But when researchers organize participants into groups paying

989 attention to their cultural backgrounds [31], and empower participants to decide which artifacts to create, they may
990 receive, in return, ecologically valid results that can be compared within and between groups to better understand
991 creative and cultural divergences. As demonstrated in section 3.3, evaluations that enact artist support networks may
992 result in participants pushing themselves to create more polished artifacts. In contrast to Phraselette’s evaluation [11],
993 which was also situated in existing sites of creative practice (i.e. poetry writing), ArtKrit evaluation’s increased Discord
994 activity potentially may have been due to the accountability of not wanting to let other participants, or the researchers,
995 down.
996

997
998 Finally, if researcher-artists also participate in the study, in addition to setting norms, they too can experience
999 personal well-being benefits. For instance, Researcher 1 learned about Giuseppe Castiglione (also known as Lang
1000 Shining), a 17th century Italian Jesuit painter/Chinese missionary who blended East-West styles, when their group was
1001 deciding which image to pick for Week 3. Finding out about an ancient painter who drew Qianlong Emperors as if
1002 they were Napoleon felt meaningful to their personal artistic practice, even if the group did not decide to do his piece.
1003 Every member of the research team who participated felt grateful to have external accountability to draw again: even if
1004 the study had not resulted in any interesting insights, at least the research team had an excuse to give time and care
1005 to their neglected artistic practices. We hope that CST evaluations can be an opportunity for researchers to use their
1006 power to bring together participants in community, orchestrating meaningful opportunities for artistic engagement.
1007
1008

1009 1010 **4.3 Limitations & Future Work**

1011
1012 Though we argue that CST evaluations can be opportunities for meaningful creative engagement, they are still research
1013 studies. A research study has its own normative ground. It fails to capture other aspects of artistic milieu that are
1014 important in artist support networks, such as different audiences on social media, while it also enacts new norms, such
1015 as how institutional affiliation may cause participants to behave in correct, expected ways. Though the researchers were
1016 clear in the on-boarding script that participants may choose to use or not use ArtKrit however they want in accordance
1017 to their existing practices, more work is to be done about the refusal of not just tools, but also norms. Even if researchers
1018 intentionally set norms they believe to be positive for participants, how can they create a space for participants refuse to
1019 engage with such norms without being socially ostracized by their peers? How can researchers balance the idiosyncratic
1020 preferences, processes, and ethics of individual artists with their desire to remain in community?
1021

1022
1023 We did not attempt to isolate independent variables, as our evaluation was not a controlled study but a technology
1024 probe toward generating a wholistic view of how relationships with technology change temporally and socially. However,
1025 more specific causal conclusions may strengthen this and future studies. For instance, when a participant changed
1026 their practice, was it more due to fellow participants’ behaviors, or more due to their interactions with ArtKrit? How
1027 much exactly did researcher participation influence the study outcomes? We believe the value of rigorous measurement
1028 metrics in design studies lie not in justifying the usability of a tool, but in disentangling the sources of authority that
1029 govern behaviors in researcher-enacted creativity support networks and ecosystems.
1030

1031
1032 Our study design also had several limitations. Our participant sample was mostly young and entirely North American
1033 based. Our small Discord groups were purposefully grouped by shared backgrounds; future work could run small
1034 groups with participants from widely different backgrounds to see if emotional support and creative safety replicates,
1035 or if participants can influence each other with even more idiosyncratic practices. Though a deployment, our study was
1036 not fully “in-the-wild”: researchers grouped up participants, who all had to draw the same master study, often different
1037 than their usual “style,” even if the image was chosen via group consensus. While a few participants did post their
1038 pieces on social media, a vast majority of them—particularly self-described professionals—chose not to because they do
1039

not share their studies in the first place, opting instead to solely post fan art to build their social media presence. A true in-the-wild study could recruit Krita users from forums with little to no guidance on the master studies used with ArtKrit—if users even choose to draw master studies at all.

5 Related Work

While this paper detailed one tool in a toolbox of potential CST evaluations [45] in an attempt to shift the norms of how such evaluations are conducted, many other researchers have evaluated—or struggled with evaluating—CSTs.

5.1 Evaluating Creativity Support Tools

CSTs span a wide variety of goals, from supporting divergent thinking, to producing novel artifacts, to expanding the possible space of artistic expression [21]. CSTs typically can help reduce users' cognitive and temporal effort when co-creating artifacts, such as a written piece or an illustration [40]. As a result, evaluations of CSTs have primarily focused on rigorous comparisons, often using a baseline in which participants create artifacts with or without the system, and assessing user experience, task performance, or the quality of the resulting creative outputs, sometimes through expert or peer annotation of those artifacts [38, 39]. Common user evaluation approaches of CSTs utilize standardized metrics such as the Creativity Support Index (CSI) [12], NASA Task Load Index (NASA-TLX) [25], and the System Usability Scale (SUS) [10] and rely on quantitative analyses of interaction time and user engagement [46]. These approaches provide methodological rigor and allow relatively straightforward comparisons across tools or conditions, offering clear insights into immediate usability and performance benefits.

Prior work has raised concerns that artifact-centric metrics and ad-hoc, usability- or productivity-focused evaluation frameworks for CSTs fail to capture the complexity, diversity, and subjectivity of creative practice [38, 45]. Creativity is often messy, exploratory, and deeply personal, making it difficult to evaluate using standardized or outcome-oriented measures such as effectiveness or efficiency [46]. As a result, existing evaluations may overlook how CSTs support long-term creative development, evolving user behaviors and interactions, and experiences that extend beyond a single user or single session [7, 38, 46] shared there is limited work that examines how CSTs are appropriated, adapted, or abandoned over time in real-world creative practice. Our work builds on these critiques by exploring an alternative evaluation approach that prioritizes longitudinal and in-situ use and emergent creative practices over short-term performance metrics [14, 45].

5.2 Longitudinal Studies and Community-Based Evaluation Studies

To complement traditional in-lab, one-off, individual-user studies, evaluations of HCI tools can extend beyond temporal and social constraints, capturing longer-term use and community dynamics. For the temporal aspect, longitudinal research in HCI emphasizes observing technology use over extended periods to capture learning, appropriation, and evolving practices [24, 44, 45]. In real-world deployments, long-term use can reveal user breakdowns, workarounds, and evolving user needs or usage patterns that are often invisible in short-term evaluations [26, 30, 50]. Additionally, extended in-situ usage can uncover issues such as novelty effects [47, 48], learning curves [4, 42], customization and personalization [35, 43], as well as misuse, nonuse, and appropriation [5, 17, 36]. The researchers behind tools such as Para [29] and Phraselette [11] have conducted extended user studies that prioritize artistic agency and real-world creative practices; we are inspired by their methods, though they do not explore group-based evaluations.

For the social aspect, community-based studies further highlight how social interaction and learning, peer feedback, and shared norms shape how tools are used, interpreted, and valued [32]. When a new tool is introduced, community

dynamics—shaped by both past practices and emerging behaviors—evolve over time, revealing how adoption, adaptation, and collaborative use unfold [33]. Community-based system evaluation enables more nuanced analyses of interactions beyond the one-user-one-tool dynamic, closer to real-world scenarios [2, 39]. Our evaluation of ArtKrit addresses this gap by modeling a community-based, longitudinal evaluation of a creativity support tool situated within an artist support network [13] and creativity support ecosystem [1].

6 Conclusion

To understand how artists work with digital drawing creativity support tools over time and within artist communities, we conducted a three-week longitudinal evaluation of ArtKrit with nine participants organized into three communities of practice. Our study revealed that participants initially explored the tool broadly but gradually focused on features that supported their individual workflows. Quantitative data showed, over the three weeks, a 53% drop in participants' active usage time and a 60% decrease in the total number of actions taken in ArtKrit. As ArtKrit provides structured guidance through actions targeting composition, value, and color aspects of the drawing process, participants increasingly focused on color-related actions (+11%, 38%), while composition-related actions decreased (-13%, 35%), and value-related actions remained stable (27%), ultimately reaching an almost even distribution across the three categories. Participants took ArtKrit's feedback on matters of composition and value but stuck with their choices for color. Qualitative findings highlighted dynamics within artist support networks, where participants engaged in positive feedback and discussion, reporting that these interaction made them feel seen, reassured, motivated, and accountable. Participants had fun drawing in community with each other: our study contributes a methodology for future CST evaluations, demonstrating they can structured as meaningful, generative opportunities for artistic practice.

References

- [1] Shm Garanganao Almeda, Joy O Kim, and Bjoern Hartmann. 2025. Creativity Supportive Ecosystems: A Framework for Understanding Function and Disruption in Online Art Worlds. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems (CHI '25)*. Association for Computing Machinery, New York, NY, USA, Article 267, 17 pages. doi:10.1145/3706598.3713734
- [2] C. Aragon, S. Poon, A. Monroy-Hernández, and Diana Aragon. 2009. A tale of two online communities: fostering collaboration and creativity in scientists and children. In *Creativity & Cognition*. <http://dl.acm.org/citation.cfm?id=1640239>
- [3] Howard S Becker. 2023. *Art worlds: updated and expanded*. Univ of California Press.
- [4] Thomas Berker, Maren Hartmann, and Yves Punie. 2005. *Domestication of media and technology*. McGraw-Hill Education (UK).
- [5] Divyanshu Bhardwaj, Alexander Ponticello, Shreya Tomar, Adrian Dabrowski, and Katharina Krombholz. 2024. In Focus, Out of Privacy: The Wearer's Perspective on the Privacy Dilemma of Camera Glasses. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems (Honolulu, HI, USA) (CHI '24)*. Association for Computing Machinery, New York, NY, USA, Article 577, 18 pages. doi:10.1145/3613904.3642242
- [6] Kathryn Blair, Miriam Sturdee, Lindsay MacDonald Vermeulen, and Lora Oehlberg. 2022. Art is Not Research. Research is not Art.. In *Extended Abstracts of the 2022 CHI Conference on Human Factors in Computing Systems (New Orleans, LA, USA) (CHI EA '22)*. Association for Computing Machinery, New York, NY, USA, Article 4, 9 pages. doi:10.1145/3491101.3516391
- [7] Jessica Blaynee, Dr David Kreps, Dr Maria Kutar, and Dr Marie Griffiths. 2016. Collaborative HCI and UX: longitudinal diary studies as a means of uncovering barriers to digital adoption. In *Proceedings of the 30th International BCS Human Computer Interaction Conference: Fusion! (Poole, United Kingdom) (HCI '16)*. BCS Learning & Development Ltd., Swindon, GBR, Article 53, 6 pages. doi:10.14236/ewic/HCI2016.72
- [8] Susanne Bødker, Christian Dindler, Ole S Iversen, and Rachel C Smith. 2022. What is participatory design? In *Participatory design*. Springer, 5–13.
- [9] Virginia Braun and Victoria Clarke. 2006. Using thematic analysis in psychology. *Qualitative research in psychology* 3, 2 (2006), 77–101.
- [10] John Brooke. [n. d.]. SUS - A quick and dirty usability scale. ([n. d.]).
- [11] Alex Calderwood, John Joon Young Chung, Yuqian Sun, Melissa Roemmele, and Max Kreminski. 2025. Phraselette: A Poet's Procedural Palette. In *Proceedings of the 2025 ACM Designing Interactive Systems Conference (DIS '25)*. Association for Computing Machinery, New York, NY, USA, 2701–2717. doi:10.1145/3715336.3735832
- [12] Erin Cherry and Celine Latulipe. 2014. Quantifying the Creativity Support of Digital Tools through the Creativity Support Index. *ACM Trans. Comput.-Hum. Interact.* 21, 4, Article 21 (June 2014), 25 pages. doi:10.1145/2617588
- [13] John Joon Young Chung, Shiqing He, and Eytan Adar. 2022. Artist Support Networks: Implications for Future Creativity Support Tools. In *Proceedings of the 2022 ACM Designing Interactive Systems Conference (Virtual Event, Australia) (DIS '22)*. Association for Computing Machinery, New York, NY,

- 1145 USA, 232–246. doi:10.1145/3532106.3533505
- 1146 [14] Tim Coughlan and Peter Johnson. 2009. Understanding productive, structural and longitudinal interactions in the design of tools for creative
1147 activities. In *Proceedings of the seventh ACM conference on Creativity and cognition*. 155–164.
- 1148 [15] Audrey Desjardins and Aubree Ball. 2018. Revealing Tensions in Autobiographical Design in HCI. In *Proceedings of the 2018 Designing Interactive
1149 Systems Conference* (Hong Kong, China) (*DIS '18*). Association for Computing Machinery, New York, NY, USA, 753–764. doi:10.1145/3196709.3196781
- 1150 [16] Audrey Desjardins and Cayla Key. 2020. Parallels, Tangents, and Loops: Reflections on the 'Through' Part of RtD. In *Proceedings of the 2020 ACM
1151 Designing Interactive Systems Conference* (Eindhoven, Netherlands) (*DIS '20*). Association for Computing Machinery, New York, NY, USA, 2133–2147.
1152 doi:10.1145/3357236.3395586
- 1153 [17] Alan Dix. 2007. Designing for appropriation. In *Proceedings of the 21st British HCI Group Annual Conference on People and Computers: HCI...but Not
1154 as We Know It - Volume 2* (University of Lancaster, United Kingdom) (*BCS-HCI '07*). BCS Learning & Development Ltd., Swindon, GBR, 27–30.
- 1155 [18] Adam G. Emerson, Shreyosi Endow, and Cesar Torres. 2024. Shared, Shaped, and Stolen: Tracing Sites of Knowledge Transfer across Creative
1156 Communities of Practice. In *Proceedings of the 16th Conference on Creativity & Cognition* (Chicago, IL, USA) (*C&C '24*). Association for Computing
1157 Machinery, New York, NY, USA, 638–650. doi:10.1145/3635636.3656199
- 1158 [19] Thomas Erickson and Wendy A. Kellogg. 2000. Social translucence: An approach to designing systems that support social processes. *ACM
1159 Transactions on Computer-Human Interaction* 7, 1 (2000), 59–83. doi:10.1145/344949.345004
- 1160 [20] Krita Foundation. 2025. Krita (Version 5.2.9). <https://krita.org> Released on January 29, 2025. Free and open-source digital painting software..
- 1161 [21] Jonas Frich, Lindsay MacDonald Vermeulen, Christian Remy, Michael Mose Biskjaer, and Peter Dalsgaard. 2019. Mapping the Landscape of Creativity
1162 Support Tools in HCI. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery,
1163 Glasgow Scotland Uk, 1–18. doi:10.1145/3290605.3300619
- 1164 [22] Bill Gaver, Tony Dunne, and Elena Pacenti. 1999. Design: Cultural probes. *Interactions* 6, 1 (Jan. 1999), 21–29. doi:10.1145/291224.291235
- 1165 [23] William Gaver. 2012. What should we expect from research through design?. In *Proceedings of the SIGCHI Conference on Human Factors in Computing
1166 Systems* (Austin, Texas, USA) (*CHI '12*). Association for Computing Machinery, New York, NY, USA, 937–946. doi:10.1145/2207676.2208538
- 1167 [24] Saul Greenberg and Bill Buxton. 2008. Usability evaluation considered harmful (some of the time). In *Proceedings of the SIGCHI Conference on
1168 Human Factors in Computing Systems* (Florence, Italy) (*CHI '08*). Association for Computing Machinery, New York, NY, USA, 111–120. doi:10.1145/
1169 1357054.1357074
- 1170 [25] Sandra G Hart and Lowell E Staveland. 1988. Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. In
1171 *Advances in psychology*. Vol. 52. Elsevier, 139–183.
- 1172 [26] A. Hertzmann. 2023. The Curse of Performative User Studies. *IEEE Computer Graphics and Applications* 43, 06 (nov 2023), 112–116. doi:10.1109/
1173 MCG.2023.3315759
- 1174 [27] Tom Hewett, Mary Czerwinski, Michael Terry, Jay Nunamaker, Linda Candy, Bill Kules, and Elisabeth Sylan. 2005. Creativity support tool
1175 evaluation methods and metrics. *Creativity Support Tools* (2005), 10–24.
- 1176 [28] Hilary Hutchinson, Wendy Mackay, Bo Westerlund, Benjamin B. Bederson, Allison Druin, Catherine Plaisant, Michel Beaudouin-Lafon, Stéphane
1177 Conversy, Helen Evans, Heiko Hansen, Nicolas Roussel, and Björn Eiderbäck. 2003. Technology probes: inspiring design for and with families. In
1178 *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Ft. Lauderdale, Florida, USA) (*CHI '03*). Association for Computing
1179 Machinery, New York, NY, USA, 17–24. doi:10.1145/642611.642616
- 1180 [29] Jennifer Jacobs, Sumit Gogia, Radomir Mundefinedch, and Joel R. Brandt. 2017. Supporting Expressive Procedural Art Creation through Direct
1181 Manipulation. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems* (Denver, Colorado, USA) (*CHI '17*). Association for
1182 Computing Machinery, New York, NY, USA, 6330–6341. doi:10.1145/3025453.3025927
- 1183 [30] Mohit Jain and Ravin Balakrishnan. 2012. User learning and performance with bezel menus. In *Proceedings of the SIGCHI Conference on Human
1184 Factors in Computing Systems*. 2221–2230.
- 1185 [31] Jun Kato, Kenta Hara, and Nao Hirasawa. 2024. Griffith: A Storyboarding Tool Designed with Japanese Animation Professionals. In *Proceedings of
1186 the 2024 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) (*CHI '24*). Association for Computing Machinery, New York,
1187 NY, USA, Article 233, 14 pages. doi:10.1145/3613904.3642121
- 1188 [32] Joy Kim, Maneesh Agrawala, and Michael S. Bernstein. 2017. Mosaic: Designing Online Creative Communities for Sharing Works-in-Progress.
1189 In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing* (Portland, Oregon, USA) (*CSCW '17*).
1190 Association for Computing Machinery, New York, NY, USA, 246–258. doi:10.1145/2998181.2998195
- 1191 [33] Obada Kraishan. 2025. The AI Attribution Paradox: Transparency as Social Strategy in Open-Source Software Development. doi:10.48550/arXiv.
1192 2512.00867 arXiv:2512.00867 [cs].
- 1193 [34] Max Kreminski and Michael Mateas. 2021. Reflective Creators. In *Proceedings of the 2021 International Conference on Computational Creativity*.
1194 International Conference on Computational Creativity, Online, 309–318. https://mkremins.github.io/publications/ReflectiveCreators_ICCC2021.pdf
- 1195 [35] Min Kyung Lee, Jodi Forlizzi, Sara Kiesler, Paul Rybski, John Antanitis, and Sarun Savetsila. 2012. Personalization in HRI: a longitudinal field
1196 experiment. In *Proceedings of the Seventh Annual ACM/IEEE International Conference on Human-Robot Interaction* (Boston, Massachusetts, USA) (*HRI
1197 '12*). Association for Computing Machinery, New York, NY, USA, 319–326. doi:10.1145/2157689.2157804
- 1198 [36] Isabel Li, Ace Chen, Eric Rawn, Shm Garanganao Almeda, Bjoern Hartmann, and Jingyi. Li. 2025. Reimagining Misuse as Creative Practice:
1199 Impressions and Implications of Usage Norms on Digital Artists. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*
1200 (Yokohama, Japan) (*CHI '25*). Association for Computing Machinery, New York, NY, USA. doi:10.1145/3706598.3714068

- 1197 [37] Jingyi Li, Eric Rawn, Jacob Ritchie, Jasper Tran O'Leary, and Sean Follmer. 2023. Beyond the Artifact: Power as a Lens for Creativity Support Tools.
1198 In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology* (San Francisco, CA, USA) (*UIST '23*). Association for
1199 Computing Machinery, New York, NY, USA, Article 47, 15 pages. doi:10.1145/3586183.3606831
- 1200 [38] Tao Long, Katy Ilonka Gero, and Lydia B Chilton. 2024. Not Just Novelty: A Longitudinal Study on Utility and Customization of an AI Workflow. In
1201 *Proceedings of the 2024 ACM Designing Interactive Systems Conference* (Copenhagen, Denmark) (*DIS '24*). Association for Computing Machinery,
1202 New York, NY, USA, 782–803. doi:10.1145/3643834.3661587
- 1203 [39] Tao Long, Sitong Wang, Émilie Fabre, Tony Wang, Anup Sathya, Jason Wu, Savvas Dimitrios Petridis, Ding Li, Tuhin Chakrabarty, Yue Jiang, Jingyi
1204 Li, Tiffany Tseng, Ken Nakagaki, Qian Yang, Nikolas Martelaro, Jeffrey V Nickerson, and Lydia B Chilton. 2025. Facilitating Longitudinal Interaction
1205 Studies of AI Systems. In *Adjunct Proceedings of the 38th Annual ACM Symposium on User Interface Software and Technology* (*UIST Adjunct '25*).
1206 Association for Computing Machinery, New York, NY, USA, Article 13, 5 pages. doi:10.1145/3746058.3758469
- 1207 [40] Tao Long, Dorothy Zhang, Grace Li, Batool Taraif, Samia Menon, Kynneddy Simone Smith, Sitong Wang, Katy Ilonka Gero, and Lydia B. Chilton.
1208 2023. Tweetorial Hooks: Generative AI Tools to Motivate Science on Social Media. doi:10.48550/arXiv.2305.12265 arXiv:2305.12265 [cs].
- 1209 [41] Jiaju Ma, Chau Vu, Asya Lyubavina, Catherine Liu, and Jingyi Li. 2025. Computational Scaffolding of Composition, Value, and Color for Disciplined
1210 Drawing. In *Proceedings of the 38th Annual ACM Symposium on User Interface Software and Technology* (*UIST '25*). Association for Computing
1211 Machinery, New York, NY, USA, Article 161, 15 pages. doi:10.1145/3746059.3747605
- 1212 [42] Thomas P Moran. 2002. Everyday adaptive design. In *Proceedings of the 4th conference on Designing interactive systems: processes, practices, methods,
1213 and techniques* (London England, 2002-06-25). ACM, 13–14. doi:10.1145/778712.778715
- 1214 [43] Michael Muller, Katja Neureiter, Nervo Verdezoto, Alina Krischkowsky, Anna Maria Al Zubaidi-Polli, and Manfred Tscheligi. 2016. Collaborative
1215 Appropriation: How Couples, Teams, Groups and Communities Adapt and Adopt Technologies. In *Proceedings of the 19th ACM Conference on
1216 Computer Supported Cooperative Work and Social Computing Companion* (*CSCW '16 Companion*). Association for Computing Machinery, New York,
1217 NY, USA, 473–480. doi:10.1145/2818052.2855508
- 1218 [44] Peter Axel Nielsen. 2021. Longitudinal Studies in Information Systems. In *Advances in Longitudinal HCI Research*, Evangelos Karapanos, Jens
1219 Gerken, Jesper Kjeldskov, and Mikael B. Skov (Eds.). Springer International Publishing, 41–56. doi:10.1007/978-3-030-67322-2_3
- 1220 [45] Christian Remy, Lindsay MacDonald Vermeulen, Jonas Frich, Michael Mose Biskjaer, and Peter Dalsgaard. 2020. Evaluating Creativity Support
1221 Tools in HCI Research. In *Proceedings of the 2020 ACM Designing Interactive Systems Conference* (Eindhoven, Netherlands) (*DIS '20*). Association for
1222 Computing Machinery, New York, NY, USA, 457–476. doi:10.1145/3357236.3395474
- 1223 [46] Samuel Rhys Cox, Helena Bøjer Djernæs, and Niels van Berkel. 2025. Beyond Productivity: Rethinking the Impact of Creativity Support Tools.
1224 In *Proceedings of the 2025 Conference on Creativity and Cognition* ('25). Association for Computing Machinery, New York, NY, USA, 735–749.
1225 doi:10.1145/3698061.3726924
- 1226 [47] Soheil Sadeghi, Somit Gupta, Stefan Gramatovici, Jiannan Lu, Hao Ai, and Ruhan Zhang. 2022. Novelty and Primacy: A Long-Term Estimator for
1227 Online Experiments. *Technometrics* 64, 4 (2022), 524–534. doi:10.1080/00401706.2022.2124309 arXiv:https://doi.org/10.1080/00401706.2022.2124309
- 1228 [48] Grace Shin, Yuanyuan Feng, Mohammad Hossein Jarrahi, and Nicci Gafinowitz. 2018. Beyond novelty effect: a mixed-methods explo-
1229 ration into the motivation for long-term activity tracker use. *JAMIA Open* 2, 1 (2018), 62–72. doi:10.1093/jamiaopen/ooy048 _eprint:
1230 https://academic.oup.com/jamiaopen/article-pdf/2/1/62/32298485/ooy048.pdf.
- 1231 [49] Ben Shneiderman. 2000. Creating creativity: user interfaces for supporting innovation. *ACM Trans. Comput.-Hum. Interact.* 7, 1 (March 2000),
1232 114–138. doi:10.1145/344949.345077
- 1233 [50] Adam J Sporka, Torsten Felzer, Sri H Kurniawan, Ondřej Poláček, Paul Haiduk, and I Scott MacKenzie. 2011. Chanti: Predictive text entry using
1234 non-verbal vocal input. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. 2463–2472.
- 1235 [51] Martin Tricaud and Michel Beaudouin-Lafon. 2023. Revisiting Creative Behaviour as an Epistemic Process: Lessons from 12 Computational Artists
1236 and Designers. In *Proceedings of the 35th Australian Computer-Human Interaction Conference* (*OzCHI '23*). Association for Computing Machinery,
1237 New York, NY, USA, 175–190. doi:10.1145/3638380.3638395
- 1238 [52] Etienne Wenger. 1998. *Communities of Practice: Learning, Meaning, and Identity*. Cambridge University Press, Cambridge, UK.
- 1239 [53] Haoqi Zhang. 2024. Searching for the Non-Consequential: Dialectical Activities in HCI and the Limits of Computers. In *Proceedings of the 2024 CHI
1240 Conference on Human Factors in Computing Systems*. Association for Computing Machinery, New York, NY, USA, 1–13. doi:10.1145/3613904.3641945