This manuscript is a preprint and has not been peer-reviewed.

Artificial Intelligence Enhances Human Creativity Through Real-Time Evaluative Feedback

Pier-Luc de Chantal [a†]
de_chantal.pier-luc@uqam.ca

Roger E. Beaty [b†]
rebeaty@psu.edu

Antonio Laverghetta Jr [b]
aml7990@psu.edu

Jimmy Pronchick [b]
jzp6211@psu.edu

John D. Patterson [b]
jdp5843@psu.edu

Peter Organisciak [c]
peter.organisciak@du.edu

Katarzyna Potega vel Żabik [d]
katarzyna.potega@kopernik.org.pl

Baptiste Barbot [e]
baptiste.barbot@uclouvain.be

Maciej Karwowski [f]
maciej.karwowski@uwr.edu.pl

† These authors contributed equally to this work.

[a] Université du Québec à Montréal
405 Sainte-Catherine Street East, Montreal, QC H2L 2C4, Canada

[b] Pennsylvania State University
201 Old Main, University Park, PA 16802, United States

[c] University of Denver
2199 South University Boulevard, Denver, CO 80210, United States

[d] Copernicus Science Centre, Warsaw, Poland.
20 Wybrzeże Kościuszkowskie, 00-390, Warsaw, Poland

[e] UCLouvain, Belgium.
1 Place de l'Université, 1348 Ottignies-Louvain-la-Neuve, Belgium

[f] University of Wroclaw, Poland.
1 plac Uniwersytecki, 50-137 Wrocław, Poland

Corresponding author: Pier-Luc de Chantal
de_chantal.pier-luc@uqam.ca
Département de Psychologie,
Université du Québec à Montréal,
C.P. 8888 succursale Centre-ville,
Montréal (Québec), H3C 3P8

Author Note

Conflict of interest: The authors declare no conflicts of interest.

Funding: This research was supported by the Social Sciences and Humanities Research Council of Canada (grant number 430-2022-00931); the National Science Foundation (grant numbers DRL-1920653, DRL-240078, and DUE-2155070); and National Science Centre of Poland (grant number 2022/45/B/HS6/00372).

Ethical statement: All procedures were performed in compliance with relevant laws and institutional guidelines and have been approved by the appropriate institutional committees (2022-4231-UQAM and 2023/GTAZY-UWr). The privacy rights of human subjects have been observed and informed consent was obtained for the studies.

Open access statement: The hypotheses, methods, and analysis plan for Study 1 were preregistered prior to data collection (https://aspredicted.org/bgmj-yrm4.pdf). All materials, data, and analysis scripts for both studies are publicly available on the Open Science Framework (https://osf.io/6k52d/?view_only=398df4fd924440ee9387af4f6434d422).

Author contributions: Conceptualization: REB, PLDC, MK; Formal analysis: PLDC, MK, AL; Investigation: REB, PLDC, KPVZ, MK; Methodology: REB, PLDC, MK, PO; Software: PLDC, JDP, PO; Visualization: PLDC; Supervision: REB, MK; Writing—original draft: REB, PLDC, MK; Writing—review & editing: AL, JP, BB, PO

Abstract

The integration of artificial intelligence (AI) into creative work continues to expand, yet its impact on human creativity itself—beyond simply providing ideas—remains uncertain. We reposition AI's role from idea generator to idea evaluator, using trained models to provide real-time feedback on human-generated ideas. Across two studies—a preregistered online experiment involving individuals with varying levels of expertise (N = 554) and a large-scale naturalistic experiment during a year-long museum exhibit (N = 36,198)—participants generated solutions to real-world problems or created visual sketches. AI feedback significantly improved participant originality in both verbal and visual creative domains. Mediation analyses revealed these gains were partly driven by changes in individuals' self-evaluation of their own originality, implying a key role for metacognition—the ability to monitor, control and regulate one's thinking. These findings suggest that AI's potential extends beyond generation to include idea evaluation, helping humans assess and refine their ideas through real-time feedback.

Keywords: *Artificial Intelligence; Creativity; Feedback; Metacognition*

Artificial Intelligence Enhances Creativity Through Real-Time Evaluative Feedback

The rise of generative AI tools has led to a surge in human-AI collaborative creativity, with people increasingly using AI to generate ideas for creative work (Bick, Blandin et al., 2024; Rafner, Beaty et al., 2023). Recent studies indicate that AI systems can produce novel and useful outputs—compared to people working without AI assistance (Doshi & Hauser, 2024; Haase & Hanel, 2023; Hubert, Awa et al., 2024; Urban, Děchtěrenko et al., 2024; Zhou & Lee, 2024)—by prompting these systems to produce text, images, music, and videos. Yet, a crucial question remains: Can AI actually help people *learn* to be more creative themselves, beyond merely prompting it for ideas? This question is especially pressing given emerging evidence that AI not only produces less diverse ideas than humans (Anderson, Shah et al., 2024; Moon, Green et al., 2024), but may actually diminish human creativity when people become dependent on it (Liu, Zhou et al., 2024). Understanding how people learn from AI interaction is therefore critical for enhancing rather than undermining human creativity.

Here, we introduce an approach to human-AI co-creativity that shifts the focus to AI evaluation and feedback, with models trained to simulate human evaluations of creativity. Our approach allows a previously unfeasible form of *real-time feedback* intervention for people generating their own ideas. Feedback is known to benefit learning and performance across many different tasks (Hattie & Timperley, 2016; Kluger & DeNisi, 1996; Wisniewski, Zierer et al., 2019), but it is difficult for human evaluators to provide reliable feedback at scale. We propose that AI feedback can support human creativity and metacognition—the ability to monitor, control, and evaluate cognitive processes relevant for creativity (Lebuda & Benedek, 2023; Puente-Diaz, 2023; K. Urban & M. Urban, 2023; M. Urban & K. Urban, 2023). In two experiments across diverse creative domains—generating solutions to real-world problems and sketching

drawings—we investigate whether real-time AI feedback can augment human creativity and improve people's ability to accurately evaluate their own ideas.

Models of creative thinking typically define two main stages: idea generation and idea evaluation (Sowden, Pringle et al., 2014; Volle, 2018; Ward, Smith et al., 1998). The advent of generative AI, particularly large language models (LLMs), has sparked widespread interest in their potential to enhance idea generation, building on a long history of research on human-AI creativity that predates LLMs (Boden, 1998). LLMs are neural networks trained on vast amounts of data, allowing them to perform a range of tasks, from answering questions to generating code to composing poetry.

The use of LLMs in creative processes has dramatically increased in recent years, with several studies assessing how LLMs impact human idea generation (Doshi & Hauser, 2024; Moon et al., 2024; Rafner et al., 2023; Urban et al., 2024; Zhou & Lee, 2024). On the one hand, because LLMs can quickly generate many ideas, they can augment human idea generation on some tasks that benefit from brainstorming. However, emerging evidence also points to potential limitations of LLM creativity, such as homogenization of outputs, resulting in reduced idea diversity (Anderson et al., 2024; Moon et al., 2024) and even reduced human idea generation when LLM access is removed (Liu et al., 2024)—suggesting minimal learning and potential de-skilling of human creativity. Furthermore, while studies frequently claim that co-creating with AI enhances human creativity, it remains unclear whether this truly occurs or if humans simply prompt LLMs and copy their outputs, raising questions about the nature of creativity enhancement with AI. As LLMs become more advanced and widely used, it is crucial to examine how they impact human creativity and to develop strategies for leveraging their strengths while mitigating potential drawbacks.

In addition to idea generation, LLMs have been used for idea evaluation, particularly for automatically evaluating responses to creativity tasks (DiStefano, Patterson et al., 2024; Jaschek, von Thienen et al., 2023; Luchini, Maliakkal et al., 2025; Organisciak, Acar et al., 2023). For example, researchers have shown strong performance in predicting human originality ratings on text-based tasks, such as generating creative uses for everyday objects, by fine-tuning pretrained language models like T5 and GPT-3 (Organisciak et al., 2023). This approach has been successfully applied to similar tasks with elementary school children, demonstrating its scalability to diverse contexts (Acar, Dumas et al., 2024). During fine-tuning, the pretrained model is adjusted for creativity ratings on specific tasks by training it on thousands of human ratings. Remarkably, GPT-4 was able to predict human creativity ratings with “few-shot” prompting, where a general model, without additional fine-tuning, is given instructions and 5-10 examples (Organisciak et al., 2023). In the visual domain, Patterson et al. trained a deep convolutional neural network called Automated Drawing Assessment (AuDrA) on thousands of line drawings and human originality ratings (Patterson, Barbot et al., 2024). The model accurately predicted the originality of new drawings (r = .80), beyond merely the drawing's complexity (i.e., ink on the page), suggesting the model accounted for nuanced representations of visual creativity (Acar, Organisciak et al., 2024; Cropley & Marrone, 2022).

Given that AI models can accurately predict creativity ratings, we evaluate a use of those predictions for *real-time feedback* while people are generating ideas. Specifically, we pursue the question: Can AI enhance human creativity by helping individuals to evaluate their own ideas? Feedback plays a critical role in learning, with extensive research showing feedback benefits for educational and professional achievement (Hattie & Timperley, 2016; Kluger & DeNisi, 1996; Wisniewski et al., 2019).

However, the link between feedback and creativity specifically has been inconsistent—while some studies show enhanced creative performance with feedback (Ezzat, Camarda et al., 2017; Gong & Zhang, 2017), others find null or even negative effects (Li, Zhang et al., 2024; Sijbom, Anseel et al., 2017).

To our knowledge, only one set of studies have investigated automated feedback on creative tasks (de Chantal, Houde-Labrecque et al., 2024; de Chantal & Organisciak, 2023), albeit using semantic distance, an earlier computational approach with a weaker correlation to human creativity ratings ($r \approx .30$), finding it helped people produce more uncommon responses to the task. Notably, the authors suggested that feedback may have influenced creative metacognition—the ability to monitor, control, and evaluate cognitive processes relevant for creativity (Lebuda & Benedek, 2023). Accurate metacognitive judgements are critical for effective idea evaluation (Puente-Diaz, 2023), yet people often struggle with evaluating their own ideas (Puente-Díaz, Cavazos-Arroyo et al., 2021), frequently exhibiting patterns of over- and under-estimation of creativity (Pesout & Nietfeld, 2021; Sidi, Torgovitsky et al., 2020; Urban & Urban, 2021). Metacognitive accuracy may therefore be a promising target for enhancing creative performance. However, receiving feedback on semantic distance did not improve accuracy (de Chantal et al., 2024), suggesting the possibility that AI feedback, more in tune with human judgments, may yield better results.

We conducted two complementary studies examining the impact of real-time AI feedback on creativity: a preregistered online experiment (N = 554) where participants generated solutions to real-world design problems, and a naturalistic museum experiment (N = 36,198) where participants sketched line drawings. Both studies delivered AI-generated originality feedback using custom-trained models and measured changes in participants' idea generation and self-evaluation accuracy relative to control

conditions. This approach enabled us to assess the impact of AI feedback across both controlled and naturalistic settings, as well as across verbal and visual modalities, thereby ensuring robust and ecologically valid findings. Given their convergence in experimental design and analysis, results from the two studies are presented together.

Method for Study 1

Participants

A total of 629 adults, aged between 18 and 71 years (*M* age = 29.80, *SD* = 10.56; 393 men, 232 women, 4 unspecified), were recruited online through the Prolific academic database. Participants were English-speaking volunteers from the United States, Canada, and the United Kingdom, 62% of whom indicated experience in an engineering field on Prolific (*n* = 388) to test whether domain expertise impacts feedback receptivity and overall performance on the design problems task. Detailed sociodemographic characteristics by engineering status (engineers, non-engineers) can be found in the Supplementary materials. As per the study's preregistration, the minimal sample size was determined through an a priori power analysis for an ANOVA with a two-level between-participants factor, using G*Power (version 3.1.9.7). We hypothesized a moderate effect of condition (*d* = 0.57) on originality based on a previous study (de Chantal & Organisciak, 2023), with alpha set at .05 and power at .90. This resulted in a minimum sample size of 210 participants. With a sample size of over 600 participants, including both engineers and non-engineers, Study 1 has sufficient power to detect the predicted effect size.

Materials

Design problems. Participants completed three design problems (Patterson, Pronchick et al., 2024), each asking them to generate as many solutions as possible to

specific problems: improve access to clean water in remote areas (hereafter “water”), assist people with memory impairments remember important tasks (“memory”), and improve the safety of pedestrian crossings (“pedestrian”). In an unanalyzed practice phase, participants proposed three solutions for making remote learning more engaging and effective. In each trial, participants were instructed to “be creative”, with creative ideas described as novel (original, unique, and innovative) and effective (practical, efficient, and feasible). During each trial, participants rated their creative self-efficacy before undertaking the task, the originality of their responses during idea generation, and the perceived difficulty of the task upon completion, all on a 10-point horizontal scale. Only self-ratings of originality were analyzed in this study. After submitting each idea, participants rated its creativity on a scale from 1 "poorly creative" to 10 "very creative", with 5 corresponding to "somewhat creative." Time-on-task was set to 6 minutes to accommodate the additional time needed for self-rating and feedback. Participants had to wait until the end of the allotted time if they ran out of ideas, with a maximum of 10 ideas allowed per trial.

Development of AI feedback. Each idea provided by participants in the design tasks was scored for originality in real-time by a fine-tuned language model, GPT-3, via an Application Programming Interface (API). The model was trained specifically for this study using an existing sample of design ideas (N = 7,705) and expert human ratings (N = 20 raters; see Supplementary materials for model training procedure). This model produced originality scores ranging from 10 to 50 (Organisciak et al., 2023), and showed strong correlations with human ratings for unseen responses in a test dataset (r = .80, p < .001; see Figure 1). The feedback scores displayed to participants were automatically scaled to 1-10 using a min-max normalization method, with the prior minimum and maximum values corresponding to the empirical minimum and maximum values of the

training dataset (15.76 and 38.24, respectively). Once the API returned the originality score for a submitted idea, this value was displayed in a half-donut graph with three possible colors: not original (gray; 3.99 or less), original (yellow; 4.00 to 6.99), and highly original (green; 7.00 or more). The specific idea submitted, along with its rounded-down originality score, was shown below the graph. Participants were encouraged to produce ideas with originality scores above 4, as indicated by the feedback.

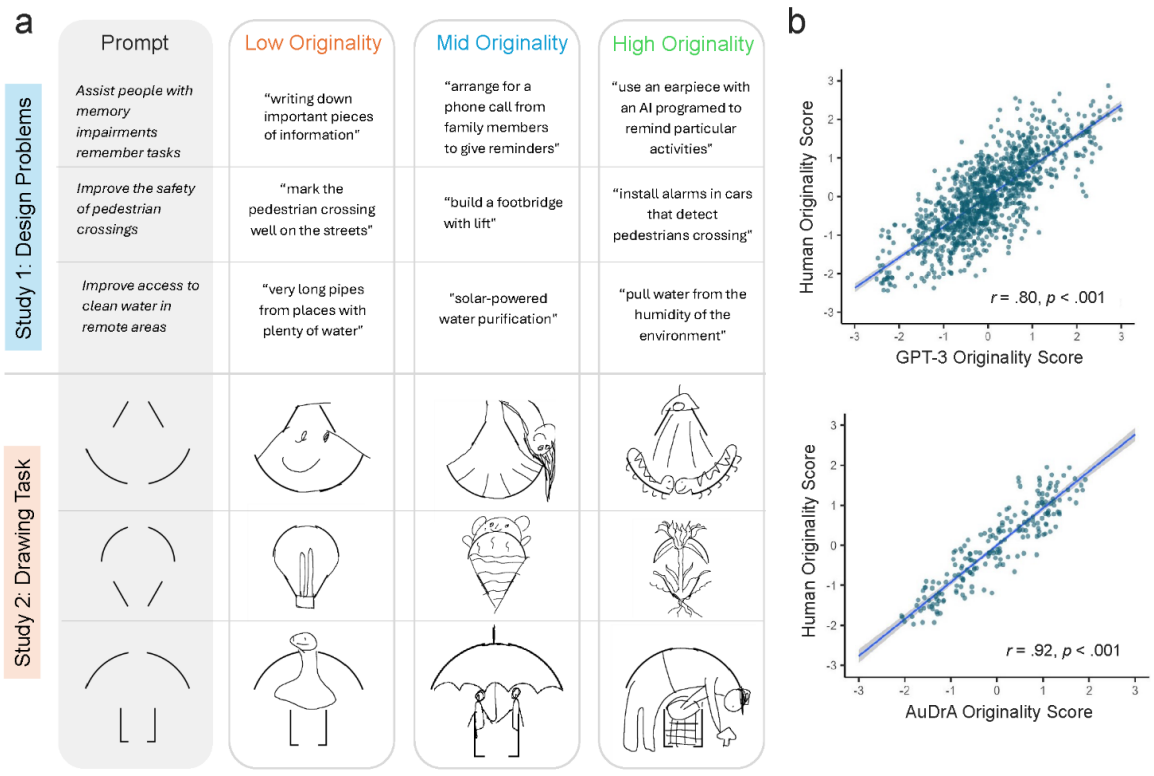


Figure 1. Representative responses and AI model validation. (a) Example design problem with responses from Study 1 (top) and drawing stimulus with completions from Study 2 (bottom), rated by AI systems as low, medium, or high originality. **(b)** Correlations between human and AI originality ratings for the GPT-3 model (top) and the AuDrA model (bottom). Shaded areas show 95% confidence intervals, and axes display standardized scores.

Procedure

The Research Ethics Board at the first author's university approved the project. The experiment was administered online using the Prolific database, with the Qualtrics survey available in open access. Figure 2 displays the experimental design. Condition was a between-participants factor with three levels (continuous feedback, intermediate feedback, no feedback). Participants in the intermediate condition received feedback during training but not during the trials, while participants in the control condition received no feedback at any stage. In all three conditions, participants received repeated "be-creative" prompts after submitting each idea (i.e., "keep in mind to generate creative solutions"), consistent with best practices in creativity assessment.

Participants first completed a practice phase to familiarize themselves with the experiment format. All three design problems were then administered in a random sequence. A total of 208 participants received feedback during both practice and trials (continuous feedback), 214 received feedback only during practice (intermediate feedback), and 208 received no feedback (control). On each trial, participants provided judgments of creative self-efficacy beforehand and rated the originality of their ideas after submitting a response. After completing the creative thinking tasks, participants answered questions about their familiarity with the problems and their appreciation of the feedback intervention (de Chantal & Organisciak, 2023). Finally, they provided basic demographic information (sex at birth, age, level of education, and experience in engineering). The study took 33.62 minutes on average, and participants were paid approximately 4£.



Figure 2. Study design overview. Study 1 (N = 554): Participants completed design problems under one of three feedback conditions—continuous feedback (practice and test phases), intermediate feedback (practice phase only), or no feedback. Study 2 (N = 36,198): Participants completed drawing tasks with either immediate feedback (after the first drawing) or delayed feedback (after two drawings). In both studies, participants generated a solution/drawing and rated its originality. Blue boxes indicate AI feedback delivery and arrows represent experimental flow.

Method for Study 2

Participants

A total of 37,823 people (*M* age = 21.95, *SD* = 15.00; 14,777 men, 18,974 women, 1,669 non-binary, 2,403 unspecified) participated in the study during their visit to the exhibit in Poland over a period of 12 months (June 2023 to June 2024).

Materials

Drawing tasks. Participants completed a drawing task, asking them to finish one of six simple incomplete shapes, representing 6 trials (Barbot, 2018). This mirrors the

Picture Completion task within the figural version of the Torrance Tests of Creative Thinking (Torrance, 1972). In each trial, participants were instructed to complete the unfinished shapes to create something new and interesting. Prior to the task, participants assessed their creative self-efficacy. Following each trial, they rated the originality of their submitted ideas using a vertical 10-point scale. The scale ranged from "poorly creative" at the bottom to "very creative" at the top, with "somewhat creative" at the midpoint. There were no time constraints for completing the drawings, and participants were free to attempt new drawings as often as they wished, using the same or different shapes.

Development of AI feedback. Each entry provided by participants in the drawing tasks was scored by a fine-tuned large language model via an application programming interface (API). The procedure used to develop the model can be found in the Supplementary materials. AI feedback was provided via the AuDrA neural network model (J. D. Patterson et al., 2024)—which strongly predicted human originality ratings (N = 141 raters) in a separate validation analysis of the AuDrA model (r = .92, p < .001; see Figure 1). Once the API returned the originality score for a submitted idea, this value was displayed on a yellow vertical scale ranging from “poorly creative” at the bottom to “very creative” at the top, with “somewhat creative” marked at the midpoint. Alongside this, a green vertical scale showed the participant's self-rating, allowing them to visually compare their assessment with the feedback.

Procedure

The Research Ethics Board at the last author’s university approved the project. Recruitment took place at a science museum, from June 2023 to June 2024. The experiment was part of the Living Lab, a dedicated research area featuring various

research projects. Before beginning, participants were informed that the exhibit aimed to develop creativity, and they provided informed consent. For children, consent was obtained from both the child and their parent or guardian. Next, participants were asked to provide basic demographic information (sex at birth, age) and report whether they were alone or with others (classmates, family members) at the exhibit. To familiarize participants with the drawing task, a brief warm-up session was conducted. During this session, participants completed one of seven unfinished objects drawn from the Test of Creative Imagery Abilities (Jankowska & Karwowski, 2015), which were distinct from the stimuli used in the main experiment. After finishing their warm-up drawing, participants self-assessed its originality using an 11-point scale, with anchors labeled "poorly original" and "very original" and a midpoint labeled "somewhat original". After the warm-up, participants were randomly assigned to one of two conditions: immediate or delayed feedback.

In both conditions, participants were reminded of the goal—i.e., to create something new and interesting—and informed that their drawings would be assessed by artificial intelligence. In the delayed feedback condition, participants were explicitly informed that they would have to create two drawings before receiving AI feedback. Participants completed the drawing tasks on a touchscreen device, with no limit on the number of drawings they could attempt—most participants only completed a few drawings, while some created up to 25 (maximum). The first shape presented was randomly selected from six possible options (for examples, see Figure 1), but participants could switch to a different shape after their initial drawing (immediate feedback condition) or after the second (delayed feedback condition). Across the study, 18,998 participants were assigned to the immediate feedback condition, and 18,825 participants were assigned to the delayed feedback condition.

Results

Transparency and openness

The hypotheses, design, sample size, and analysis plan for Study 1 were registered prior to data collection (https://aspredicted.org/bgmj-yrm4.pdf). Notable deviations include: (a) the addition of an intermediate condition to allow a more precise test of the feedback effect; (b) the use of a correlational approach to analyze evaluation accuracy, replacing the preregistered squared-difference score method; (c) the inclusion of both engineers and non-engineers to examine potential domain-specific effects. To account for these changes, a larger sample size was recruited. All deviations are further justified in the relevant sections. Study 2 was not preregistered. All materials, data, and analysis scripts for both studies are publicly accessible on the Open Science Framework https://osf.io/6k52d/?view_only=398df4fd924440ee9387af4f6434d422.

Data cleaning

Study 1. Following preregistration, the dataset of 8,173 entries was screened for invalid ideas. This screening was accomplished using a combination of manual review by research assistants and a transformer language model (TLM) trained to detect invalid responses. Full details on the TLM training process are described in the Supplementary materials. These invalid entries accounted for less than 1% of all ideas and were excluded from further analysis. Eleven participants were excluded for completing the study in a time that was more than three standard deviations from the average duration. Additional unregistered exclusions were applied to ensure data quality. Specifically, 56 participants were identified as having over 25% of their responses showing signs of copy-and-paste behavior, as indicated by a significant discrepancy between the number of keys pressed and the number of characters in their responses. Furthermore, 15

participants used a mobile phone or tablet, and 34 trials contained only one response and were thus excluded. The distribution of excluded entries and participants was consistent across all conditions. The final dataset comprised 554 adults aged between 18 and 71 years (*M* age = 29.82, *SD* = 10.54; 343 men, 207 women, 4 unspecified), who provided 7,174 valid responses. There were 327 engineers (114 in feedback, 106 in intermediate, 107 in control conditions) and 228 non-engineers (76 in feedback, 75 in intermediate, 77 in control conditions).

Study 2. We excluded drawings whose completion time was in the bottom 3% of the sample distribution, resulting in the deletion of 2,632 entries submitted in 6 seconds or less. This step was taken to eliminate low-quality responses. The final dataset comprised 36,198 participants (*M* age = 22.0, *SD* = 15.0; 14,089 men, 18,343 women, 1,528 non-binary and 2,238 unspecified), who provided a total of 77,142 drawings. There were 18,136 people in the immediate feedback condition and 18,062 people in the delayed feedback condition.

Analysis strategy

After cleaning the datasets, we analyzed 7,174 solutions from the online experiment (Study 1) and 77,142 drawings from the museum experiment (Study 2), applying a consistent analysis strategy across both studies. We tested our hypotheses using sequential linear mixed-effects models with different sets of predictors, controlling for baseline differences by including participants and prompts (design problem, drawing shape) as random intercepts. We also conducted exploratory analyses to examine whether the effect of feedback on originality was mediated by changes in evaluation accuracy. Given the similar study design and analysis strategies, we present the results of Study 1 and Study 2 side by side to provide a clear comparison of our findings.

In Study 1, the experimental manipulation was a three-level categorical variable (continuous, intermediate, or no feedback), whereas Study 2 employed a slightly different approach. Given the naturalistic nature of this second experiment, participant engagement varied widely, with some spending significant time on the activity while others engaged minimally. Notably, after the initial drawings, even participants in the delayed feedback condition gained significant experience with feedback, making the two experimental conditions less contrasting after the first two drawings. Therefore, we restricted the analysis to a subset of the dataset that ensures a reliable comparison between the feedback (immediate feedback) and no-feedback conditions (delayed feedback). The first drawing for a given shape was always submitted without feedback in both conditions; the second drawing followed feedback only in the immediate feedback condition, and all the drawings that followed were created with feedback in both conditions. Therefore, examining only the two first drawings per shape allows for a clearer comparison between conditions with and without feedback, while also reducing the noise caused by the exhibition setting, where participants were free to generate nearly as many drawings as they desired for as many of six shapes of their choice.

The main analyses, carried out using R 4.3.2, followed a similar approach for both studies. We examined the effect of condition and self-ratings on originality using mixed-effects models, making it possible to account for the multilevel structure of the dataset. The full scripts are available as Supplementary materials. The models were fitted using the "lmer" function from the "lme4" package 1.1-35.1 (Bates, Mächler et al., 2015). The assumptions of homoscedasticity and normality of the residuals were examined by visual inspection of plots of the residuals against the fitted values and Q-Q plots of the residuals. All models satisfied the homoscedasticity assumption but sometimes indicated the presence of outliers. We computed cluster-robust variance

estimates as a robustness test using the "clubSandwich" package 0.5.10 (Pustejovsky, 2018). As the results with robust estimates converged with the original results, these are not reported here. Originality was analyzed at the response level, which is nested within participants. Therefore, originality scores were neither aggregated to the trial level nor to the participant level. In each model, random intercepts for participant and design problem were included to control for differences across participants and problem types. The significance of the effects was assessed using the "lmerTest" package 3.1-3 (Kuznetsova, Brockhoff et al., 2017), which provides p-values for the fixed effects. We also employed Wald chi-square test, from the "car" package 3.0-13 (Fox & Weisberg, 2019), to assess the significance of variables with categorical levels, as it provides an alternative, yet simplified approach to report these effects. The best-fitting model was identified using likelihood ratio tests and post-hoc analyses were performed using the "emmeans" package 1.10.1 (Lenth, Singmann et al., 2019).

In both studies, originality was first regressed on condition, and then on response order, trial order, and their two-way interactions. As an unregistered analysis of the predictors of originality, we also computed a third model including self-ratings as a fixed effect. More specifically, we had originally planned and preregistered for Study 1 to examine evaluation accuracy using the squared difference technique, which involves subtracting originality scores from self-ratings and then squaring the difference to eliminate directional bias. Instead, our correlational analysis examines the linear relationship between self-evaluations and originality scores. The argument for preferring correlations to difference scores is twofold. First, it has recently been argued that difference scores can be biased by participants' baseline (intercept) use of the rating scale (Rietzschel, Ritter et al., 2024). For instance, individuals may systematically anchor their self-ratings to a specific value, thereby reducing the range of the scale they

362 truly use and making comparisons with reference scores, which use the whole scale,
363 more difficult and unreliable. Correlations, however, capture the relationship between
364 self-ratings and originality scores independently of baseline tendencies, making them
365 less prone to bias. In other words, they assess whether the ratings follow the same
366 covariation pattern on the scale, regardless of where participants anchor their ratings.
367 Second, using correlations allows for more parsimonious reporting, as both self-ratings
368 and condition effects can be incorporated within the same predictive model. For
369 transparency, we report the analysis using squared differences in the Supplementary
370 materials.

371 For Study 1 only, we carried out an exploratory analysis to determine how
372 including engineering status (engineers, non-engineers) impacted the model fit. While
373 engineers reported higher familiarity with the problems, $d = 0.40$, $t(552) = 4.59$, $p < .001$,
374 considering engineering status did not significantly improve model fit (AIC = 26,483,
375 $\Delta\chi^2(4) = 7.70$, $p = .10$), so all participants were pooled for further analyses.

376 Finally, we carried out an additional exploratory analysis to examine whether
377 feedback enhances originality through improved evaluation accuracy. We proceeded
378 with a two-level mediation model (responses nested within participants) using the
379 “lavaan” package in R. In Study 1, the self-rating of originality served as a mediator
380 between contrast-coded condition (control = -1, intermediate feedback = 0, continuous
381 feedback = 1) and AI-scored originality. Originality was regressed on both the condition
382 variable and self-ratings. Our primary focus was on the level 2 (person-level) indirect
383 effect - i.e., we examined whether assigning participants to one of the three conditions
384 (control, intermediate feedback, or feedback) influenced their average self-rating
385 (reflecting evaluation accuracy), which, in turn, predicted the originality of their generated
386 ideas. In Study 2, the model included condition (delayed feedback = 0, immediate

feedback = 1) as the independent variable, self-ratings of the second drawing as the mediator, and the AI-score originality of the second drawing as the dependent variable. To control for baseline performance, the self-rating and AI-score originality of the first drawing were included as covariates. This approach allowed us to assess whether condition was associated with changes in self-ratings between the first and second drawings, and whether these changes in self-ratings were associated with corresponding changes in originality.

Reliability of the originality and self-rating scores

In Study 1, originality scores provided by the API were saved as participants produced ideas, resulting in scores ranging from 1 to 10 (M = 6.34, SD = 2.11). Correlations between the originality scores of each design problem ranged from .70 to .73, all ps < .001. This showed excellent internal consistency, with a Cronbach's alpha of .87, 95% CI [.85, .89]. Participants also self-rated the originality of each idea on a scale from 1 to 10 (M = 5.92, SD = 2.04), with correlations between prompts ranging from .58 to .61, all ps < .001. Internal consistency was good (α = .81, [.78, .83]). It should be noted that the correlation patterns across problems were identical for both engineers and non-engineers. Pairwise correlations between both scores are reported in the Supplementary materials. In Study 2, originality scores ranged from 0 to 10 (M = 4.57, SD = 2.46). Correlations between the scores of each trial ranged from .25 to .36, all ps < .001. This showed good internal consistency, with a Cronbach's alpha of .72, 95% CI [.72, .73]. Participants also self-rated the originality of each of their ideas on a scale from 0 to 10 (M = 6.88, SD = 3.17), with correlations between shapes ranging from .33 to .42, all ps < .001. Again, internal consistency was good (α = .79, [.79, .80]).

Does feedback drive changes in originality?

We started by examining the effect of feedback on originality scores, which were recorded as participants produced design solutions (Study 1) and drawings (Study 2). If feedback resulted in changes in the level of originality, a significant effect of condition should be found in both studies. To test this, we first regressed originality on condition alone, providing a basis for subsequent comparative models and directly examining our primary hypothesis. As predicted, the effect of condition was a significant predictor of originality in both Study 1, Wald $\chi^2(2) = 21.37$, $p < .001$, and Study 2, Wald $\chi^2(1) = 54.37$, $p < .001$, clearly demonstrating that feedback influenced overall levels of originality (see Figure 3).

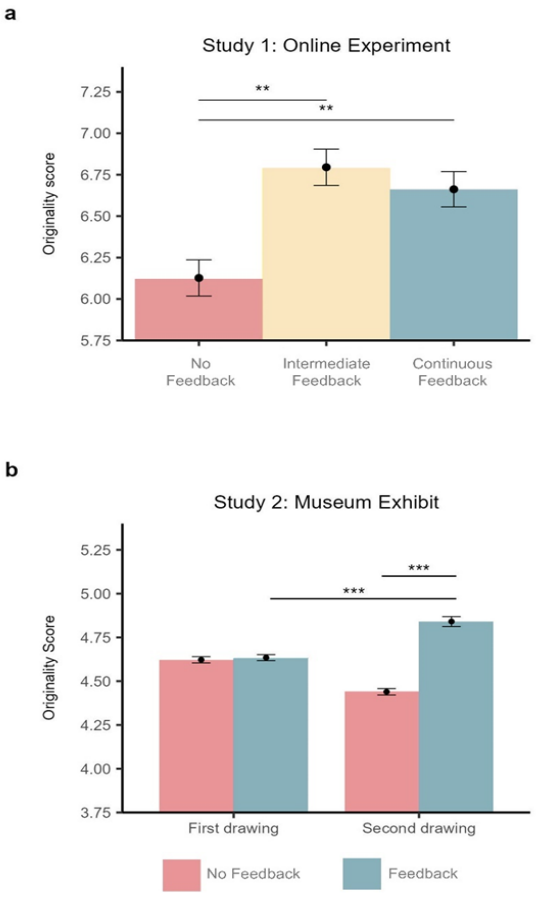


Figure 3. Estimated Marginal Mean originality scores by condition. Bars represent standard errors. The original y-axis values (1–10) were truncated to improve visualization. ** $p < .01$, *** $p < .001$. **(a)** Study 1: Originality was higher in the continuous feedback (*EMM* = 6.67, *SE* = 0.15) and intermediate conditions (*EMM* = 6.80, *SE* = 0.15) compared to control (*EMM* = 6.13, *SE* = 0.15). **(b)** Study 2: For the first drawing, there was no significant difference between the feedback (*EMM* = 4.63, *SE* = 0.02) and control conditions (*EMM* = 4.62, *SE* = 0.02). For the second drawing, originality was higher in the feedback (*EMM* = 4.84, *SE* = 0.02) than in the control condition (*EMM* = 4.44, *SE* = 0.02), with originality in the feedback condition exceeding baseline levels observed in the first drawing.

In Study 1, levels of originality were significantly higher in the continuous feedback (*EMM* = 6.67, *SE* = 0.15, 95% CI [6.34, 7.00]) than in the control condition (*EMM* = 6.13, *SE* = 0.15, 95% CI [5.79, 6.46]), $t(8.50) = -3.56$, $p < .01$, with a medium effect size ($d = -0.35$, *SE* = 0.10, 95% CI [−0.57, −0.13]). The intermediate condition also outperformed the control group (*EMM* = 6.80, *SE* = 0.15, [6.46, 7.13]), $t(8.95) = -4.36$, $p < .01$, again showing medium effect size ($d = -0.43$, *SE* = 0.10, 95% CI [−0.66, −0.21]). However, no significant differences were found between the continuous and intermediate feedback conditions ($d = 0.08$, *SE* = 0.10, 95% CI [−0.14, 0.30]), $t(3.80) = -0.86$, $p = .68$. Although this aligns with our preregistered prediction, the finding that participants in the intermediate condition—who received feedback only during practice—showed such a significant increase suggests that even minimal feedback can influence originality. These findings suggest that providing real-time AI feedback—even minimally during the practice phase alone—led to medium-sized improvements in originality in the subsequent phase. Importantly, the effect sizes are robust, as they control for

participant- and prompt-level variability, offering a conservative yet accurate estimate of the differences across conditions.

Similarly, in Study 2, participants in the feedback condition showed a significant increase in originality across the first two drawings compared to the control group ($p <$ .001), replicating Study 1. Further model estimates are available in the Supplementary materials. Post-hoc analysis in Study 2 quantified these differences, with participants in the feedback group producing drawings of higher originality (*EMM* = 4.65, *SE* = 0.02, 95% CI [4.62, 4.68]) than those in the control group (*EMM* = 4.49, *SE* = 0.02, 95% CI [4.46, 4.52]), $t(29{,}398) = -7.25$, $p < .001$. Notably, compared to Study 1, the effect size was smaller ($d = -0.08$, *SE* = 0.01, 95% CI [−0.10, −0.06]), which was expected as this comparison includes the first drawing, always produced without feedback in both conditions. The key contrast in Study 2 lies between the first (baseline) and second drawings, making response order a critically important factor to add in the predictive model.

If receiving feedback led to increased levels of originality in Study 2, we would expect this effect to be stronger for the second drawing, after receiving the AI feedback, leading to a Condition × Response Order interaction. To test this, we regressed originality on condition, response order, and trial order, as well as their interactions (Response Order × Condition and Trial Order × Condition). In both studies, trial order had the same meaning, as it referred to successive design problems (Study 1) or shapes (Study 2). However, the interpretation of response order differed. In Study 1, participants could generate up to 10 solutions per design problem, making response order a continuous variable. In Study 2, we analyzed only the first and second drawings, making response order a two-level categorical variable. This comparative model allowed us to

examine the effect of feedback while controlling for trial and response order, as well as to explore the cumulative effect across repeated solutions and drawings.

In both studies, this more complex model involving response and trial order provided a better fit to the data than the simpler one-predictor model (Study 1: $\Delta\chi^2$ (6) = 29.07, $p < .001$; Study 2: $\Delta\chi^2$ (4) = 618.55, $p < .001$; for Akaike Information Criterion values and additional model estimates, see the Supplementary materials). While the effect of condition was maintained in Study 1, Wald $\chi^2(2) = 27.35$, $p < .001$, reinforcing its effect after controlling for trial and response order, it was greatly reduced in Study 2, as it was qualified by the predicted Condition × Response Order interaction, Wald $\chi^2(2) = 140.70$, $p < .001$. For the first (baseline) drawing, estimated marginal means were nearly identical across the feedback (*EMM* = 4.63, *SE* = 0.02, 95% CI [4.60, 4.67]) and control conditions (*EMM* = 4.62, *SE* = 0.02, 95% CI [4.59, 4.66]), $t(46{,}652) = -0.50$, $p = .62$. However, the levels of originality in the second drawing were higher when preceded by feedback (*EMM* = 4.84, *SE* = 0.02, 95% CI [4.79, 4.90]) compared with no feedback (*EMM* = 4.44, *SE* = 0.02, 95% CI [4.40, 4.48]), $t(54{,}419) = -11.83$, $p < .001$, indicating a small-to-medium effect size ($d = -0.20$, *SE* = 0.02, 95% CI [−0.24, −0.17]). Taken together, these findings suggest that providing AI feedback led to improvements in originality, with robust effect sizes across both studies, although larger effects in the controlled online experiment using verbal design problems ($d = 0.43$ and 0.35) than in the naturalistic museum study employing visual sketch tasks ($d = 0.20$).

Does feedback improve evaluation accuracy?

A critical test of AI feedback is its impact on metacognitive monitoring, here operationalized as evaluation accuracy and measured using correlations between AI scores and self-ratings. We would expect that, if evaluation accuracy is related to basic

creative thinking abilities, then adding self-ratings as a fixed factor in predicting originality scores should improve model fit. More importantly, we would expect stronger relationships between self-ratings and originality scores in the feedback conditions, suggesting an interaction effect. To test these predictions in both studies, we regressed originality on condition, response variables (response order, trial order, and their interactions with condition), and added self-ratings (1-10 scale) along with its interaction with condition (Self-Ratings × Condition). For Study 2, we also included a three-way Self-Ratings × Condition × Response Order interaction to account for differences between the first (baseline) and second drawings.

In both studies, this more complex model involving self-ratings provided a better fit to the data (Study 1: $\Delta\chi^2$ (6) = 29.07, $p < .001$; Study 2: $\Delta\chi^2$ (4) = 2,081.51, $p < .001$), reinforcing the key predictive value of evaluative processes (for Akaike Information Criterion values and model estimates, see the Supplementary materials). Self-ratings significantly predicted AI originality scores, irrespective of whether participants received feedback, in both Study 1, Wald $\chi^2(1) = 602.62$, $p < .001$, and Study 2, Wald $\chi^2(1) = 689.94$, $p < .001$. An increase of one point in self-ratings was associated with a 0.41-point ($SE = 0.02$, $p < .001$) and a 0.13-point ($SE = 0.01$, $p < .001$) increase in originality scores in Study 1 and 2, respectively (see Figure 4, for additional pairwise correlations between self-ratings and originality scores by conditions, see the Supplementary materials). However, in Study 1, feedback no longer had a significant effect on originality once self-ratings were accounted for ($b = 0.12$, $SE = 0.24$, $p < .62$), suggesting that feedback influenced originality through changes in evaluation accuracy. If feedback led to significant changes in evaluation accuracy, self-ratings should be more strongly correlated with originality in the feedback than in the control conditions. Supporting this idea, self-ratings interacted with condition in Study 1, Wald $\chi^2(1) = 26.89$, $p < .001$, and

as predicted, we found a significant Self-Ratings × Condition × Response Order interaction in Study 2, Wald $\chi^2(1) = 10.95$, $p < .001$. We conducted post-hoc analyses to further explore the nature of these interactions.

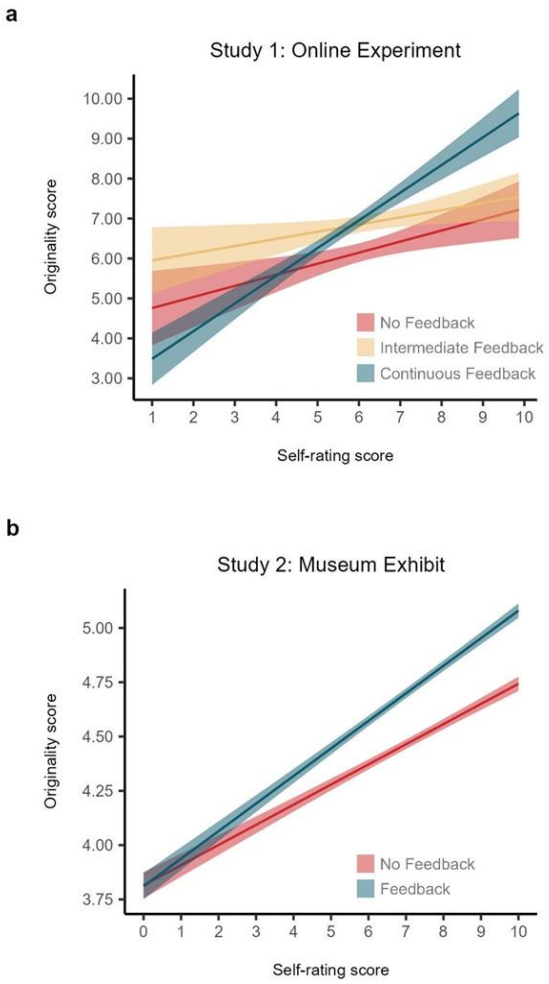


Figure 4. Fitted regression lines of self-rating and originality scores. Shaded areas surrounding the regression lines are confidence intervals. **(A)** Study 1: Correlation was stronger in the feedback condition ($r = .59$, $p < .001$) compared to the intermediate ($r = .17$, $p < .05$) and control conditions ($r = .22$, $p < .05$). **(B)** Study 2 displays the regression lines for the second drawing: Correlation was higher in the feedback ($r = .22$, $p < .001$) than in the control condition ($r = .15$, $p < .001$).

In Study 1, self-ratings consistently predicted originality scores across all conditions. However, the association was significantly stronger for participants receiving feedback (b = 0.52, SE = 0.02, 95% CI [0.49, 0.56]) compared to those in the intermediate (b = 0.40, SE = 0.02, [0.37, 0.44]), $t(7094) = -4.69$, $p < .001$, and the control conditions (b = 0.41, SE = 0.02, [0.38, 0.44]), $t(7089) = -4.45$, $p < .001$. There was no significant difference between both intermediate and control conditions, $t(7069) = 0.40$, $p = .92$ (see Figure 4). These findings clearly demonstrate that although receiving minimal feedback during practice (intermediate) led to improved idea generation (originality), it did not significantly drive better evaluation accuracy. Changes in evaluation accuracy appeared to be confined to the feedback condition, where participants received feedback not only during practice but throughout the entire creative thinking task.

The pattern was similar in Study 2, with self-ratings predicting originality similarly for the first (baseline) drawing in both the feedback (b = .12, SE = 0.01, 95% CI [.11, .14]) and control conditions (b = .13, SE = 0.01, 95% CI [.12, .14]), $t(47{,}103) = 0.59$, $p = .93$. However, it was only after receiving feedback that self-ratings became a stronger predictor of originality (b = .16, SE = 0.01, 95% CI [.14, .18]), $t(47{,}103) = -4.32$, $p < .001$, while the linear relationship remained constant in the control condition (b = .13, SE = 0.01, 95% CI [.12, .14]), $t(51{,}010) = -0.68$, $p = .90$ (for more information, see the Supplementary materials).

Do changes in self-ratings drive changes in originality?

Our previous analyses suggest that feedback influenced originality scores through changes in evaluation accuracy, though they did not provide direct evidence. To examine this possibility more specifically, we conducted an exploratory multilevel

mediation analysis. In Study 1, condition was contrast-coded (control = –1, intermediate feedback = 0, continuous feedback = 1). The between-level indirect effect was statistically significant, b = –0.10, 95% CI [–0.15, –0.04], p < .001, indicating that changes in self-ratings partially explained changes in originality. Participants in the continuous feedback condition, whose self-evaluations showed better alignment with AI scores, tended to rate their ideas as less creative compared to those in the other groups (b = –0.25, [–0.38, –0.12], p < .001), indicating that improved evaluation accuracy was driven by more modest self-ratings (for more information, see the Supplementary materials).

In Study 2, we estimated a mediation model with contrast-coded condition as the independent variable (control = 0, feedback = 1), self-ratings of the second drawing as the mediator, and the originality of the second drawing as the dependent variable. Self-ratings and originality of the first drawing were included as covariates. This mediation model allowed us to examine whether feedback was associated with changes in self-ratings (i.e., the difference in self-assessment between the second and first drawing) and whether these changes were related to differences in originality between the second and first drawing. Consistent with the results from Study 1, Study 2 revealed a statistically significant negative indirect effect, indicating that feedback led to lower yet more accurate self-assessments (b = –0.02, 95% CI [–0.03, –0.01], p < .001), which in turn predicted an increase in originality between the first and second drawing. Specifically, participants in the feedback condition showed a decrease in self-rated originality between the first and second drawing, (b = –0.14, [–0.22, –0.06], p < .001), while this change in self-ratings was positively associated with changes in originality scores, b = 0.14, 95% CI [0.13, 0.15], p < .001 (see the Supplementary materials).

Discussion

Artificial Intelligence is rapidly transforming how people engage in creative work, raising key questions about how to best leverage AI to enhance human creativity. Our results provide the first experimental evidence that AI feedback can benefit creative thinking, observed through two large-scale studies across distinct creative modalities (verbal and visual) and settings (controlled laboratory and naturalistic environment). We successfully trained AI models to simulate human evaluators, enabling real-time feedback that aligns with human perceptions of creativity—feedback that would be impractical for human evaluators to deliver at scale. Participants produced more original solutions to real-world problems (d = 0.35, Study 1) and more original drawings (d = 0.20, Study 2) when receiving AI-based numerical feedback. Importantly, improved idea generation occurred even with minimal feedback during practice in Study 1 (d = 0.43). However, enhanced evaluation accuracy required continuous feedback throughout the entire experiment. Our findings suggest that AI can serve an educational role in promoting human creativity by helping people become better at generating and evaluating their own ideas.

Importantly, the AI models used in our study were trained to mimic human judgments. In Study 1, we fine-tuned a GPT model on expert originality ratings for design problems—showing it thousands of example solutions and ratings from 40 human experts with a background in engineering and design—finding the model predicted the originality of new solutions it was not trained on with high accuracy (r = .80). Likewise, in Study 2, we used a convolutional neural network that was previously trained to evaluate creativity of line drawings (J. D. Patterson et al., 2024) we validated the model with an independent sample of 120 human raters, showing that the AI ratings strongly correlated with human creativity ratings (r = .92).

One important finding is that participants who received feedback only during practice in Study 1 produced more original solutions—a remarkable finding on its own that suggests even a single “dose” of performance feedback can boost creativity, with feedback benefits observed in both domain experts and novices with no prior training in engineering design. Such rapid learning was also observed in Study 2, where participants produced more original drawings following a single round of AI feedback. Although we did not compare against non-human aligned AI models, our results suggest that human-aligned AI can provide meaningful feedback to improve creative thinking in both verbal and visual domains.

Regarding idea evaluation, we found a different rate of feedback learning. While idea generation could be improved with a single exposure to AI feedback, idea evaluation required sustained feedback for successful learning. Participants receiving continuous feedback in Study 1 showed stronger alignment between self-ratings and AI scores compared to intermediate feedback or controls, while Study 2 evaluation accuracy improved only after feedback. Thus, while idea generation benefited from brief exposure, developing accurate evaluation may require extended practice. Critically, we found that idea evaluation accuracy partially mediated idea generation originality—AI feedback boosted idea originality in part by improving people's ability to accurately evaluate their ideas, indicating that idea evaluation is a cognitive mechanism of successful feedback learning. This finding aligns with theories proposing that accurate monitoring of creative processes supports the generation of more original ideas (Rietzschel et al., 2024), as people learn to recognize and build upon promising ideas and reject common ideas.

Taken together, our findings suggest broadening AI's role in creative work from idea generator to idea evaluator. Recasting AI for idea evaluation, and prioritizing

humans for idea generation, may help to combat emerging challenges of generative AI for human creativity, such as AI ideas being less diverse than (Anderson et al., 2024; Moon et al., 2024) and deskilling of human creativity without AI assistance (Liu et al., 2024). Although generative AI can generate original ideas on some tasks, studies show AI outputs lack diversity—AI ideas are more similar to each other than human ideas—highlighting a risk to creativity that could be mitigated by keeping humans engaged in idea generation, leveraging their unique experiences and knowledge while using human-aligned AI models for evaluation.

The benefits we observed in both laboratory and museum settings suggest AI feedback could be particularly valuable for creativity training, where human evaluators often struggle to provide consistent, timely feedback to large numbers of learners. Moreover, our finding that even brief AI feedback improved idea generation suggests this approach could be efficiently implemented in settings where sustained human feedback may be impractical. This could enhance accessibility to creativity training, allowing people with limited access to expert teachers to receive immediate feedback that mimics expert evaluation. Notably, AI feedback consisted of 1-10 originality ratings only. Future work should move beyond such quantitative-only feedback to explore qualitative feedback, leveraging generative AI to explain limitations and suggest improvements, aligned to human experts.

While our studies demonstrate that AI feedback can enhance creative performance in a single session, developing creative expertise requires sustained practice over longer time periods. The modest effect sizes observed, particularly in the museum study, reflect known ceiling effects in creative skill development—substantial improvements typically emerge through extended practice, domain knowledge acquisition, and expert guidance (Baer, 2015; Ericsson, Krampe et al., 1993). This

suggests that AI feedback systems might be most effective when integrated into broader learning contexts that include human mentorship, sustained practice opportunities, and the development of domain expertise. Future work should examine how AI feedback can complement traditional approaches to creative skill development, particularly in educational and professional settings where both immediate feedback and long-term expertise development are crucial.

References

Acar, S., Dumas, D., et al. (2024). Measuring original thinking in elementary school: Development and validation of a computational psychometric approach. *Journal of Educational Psychology*. https://doi.org/10.1037/edu0000844

Acar, S., Organisciak, P., et al. (2024). Automated scoring of figural tests of creativity with computer vision. *The Journal of Creative Behavior*. https://doi.org/10.1002/jocb.677

Anderson, B. R., Shah, J. H., et al. (2024). *Homogenization effects of large language models on human creative ideation* Creativity and Cognition,

Baer, J. (2015). The importance of domain-specific expertise in creativity. *Roeper Review*, *37*(3), 165-178. https://doi.org/10.1080/02783193.2015.1047480

Barbot, B. (2018). The Dynamics of Creative Ideation: Introducing a New Assessment Paradigm. *Frontiers in Psychology*, *9*, 2529. https://doi.org/10.3389/fpsyg.2018.02529

Bates, D., Mächler, M., et al. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1). https://doi.org/10.18637/jss.v067.i01

Bick, A., Blandin, A., et al. (2024). The rapide adoption of generative AI (Working paper no. 32966). *National Bureau of Economic Research*. https://doi.org/10.3386/w32966

Boden, M. A. (1998). Creativity and artificial intelligence. *Artificial Intelligence*, *103*(1), 347-356. https://doi.org/10.1016/S0004-3702(98)00055-1

Cropley, D. H., & Marrone, R. L. (2022). Automated scoring of figural creativity using a convolutional neural network. *Psychology of Aesthetics, Creativity, and the Arts*. https://doi.org/10.1037/aca0000510

de Chantal, P. L., Houde-Labrecque, C., et al. (2024). Investigating Lasting Effects of Real-Time Feedback on Originality and Evaluation Accuracy. *Creativity Research Journal*. https://doi.org/10.1080/10400419.2024.2441014

de Chantal, P. L., & Organisciak, P. (2023). Automated feedback and creativity: On the role of metacognitive monitoring in divergent thinking. *Psychology of Aesthetics, Creativity, and the Arts*. https://doi.org/10.1037/aca0000592

DiStefano, P. V., Patterson, J. D., et al. (2024). Automatic Scoring of Metaphor Creativity with Large Language Models. *Creativity Research Journal*, 1-15. https://doi.org/10.1080/10400419.2024.2326343

Doshi, A. R., & Hauser, O. P. (2024). Generative AI enhances individual creativity but reduces the collective diversity of novel content. *Science Advances*, *10*(28). https://doi.org/10.1126/sciadv.adn5290

Ericsson, K. A., Krampe, R. T., et al. (1993). The role of deliberate practice in the acquisition of expert performance. . *Psychological Review*, *100*(3), 363-406. https://doi.org/10.1037/0033-295X.100.3.363

Ezzat, H., Camarda, A., et al. (2017). How minimal executive feedback influences creative idea generation. *PLoS One*, *12*(6), e0180458. https://doi.org/10.1371/journal.pone.0180458

Fox, J., & Weisberg, S. (2019). *An R Companion to Applied Regression* (Third ed.). Sage.

Gong, Z., & Zhang, N. (2017). Using a Feedback Environment to Improve Creative Performance: A Dynamic Affect Perspective. *Front Psychol*, *8*, 1398. https://doi.org/10.3389/fpsyg.2017.01398

Haase, J., & Hanel, P. H. P. (2023). Artificial muses: Generative artificial intelligence chatbots have risen to human-level creativity. *Journal of Creativity*, *33*(3). https://doi.org/10.1016/j.yjoc.2023.100066

Hattie, J., & Timperley, H. (2016). The power of feedback. *Review of Educational Research*, *77*(1), 81-112. https://doi.org/10.3102/003465430298487

Hubert, K. F., Awa, K. N., et al. (2024). The current state of artificial intelligence generative language models is more creative than humans on divergent thinking tasks. *Sci Rep*, *14*(1), 3440. https://doi.org/10.1038/s41598-024-53303-w

Jankowska, D. M., & Karwowski, M. (2015). Measuring creative imagery abilities. *Frontiers in Psychology*, *6*, 1591. https://doi.org/10.3389/fpsyg.2015.01591

Jaschek, C., von Thienen, J., et al. (2023). The CollaboUse Test for Automated Creativity Measurement in Individuals and Teams: A Construct Validation Study. *Creativity Research Journal*, *35*(4), 677-697. https://doi.org/10.1080/10400419.2022.2150813

Kluger, A. N., & DeNisi, A. (1996). The effects of feedback interventions on performance: A historical review, a meta-analysis, and a preliminary feedback intervention theory. *Psychological Bulletin*, *119*, 254-284.

Kuznetsova, A., Brockhoff, P. B., et al. (2017). lmerTest Package: Tests in Linear Mixed Effects Models. *Journal of Statistical Software*, *82*(13). https://doi.org/10.18637/jss.v082.i13

Lebuda, I., & Benedek, M. (2023). A systematic framework of creative metacognition. *Physics of Life Reviews*, *46*, 161-181. https://doi.org/10.1016/j.plrev.2023.07.002

Lenth, R., Singmann, H., et al. (2019). emmeans: Estimated marginal means, aka least squares means.

Li, H., Zhang, J., et al. (2024). Exploring the effects of negative supervisory feedback on creativity among research and development personnel: challenge or threat? *Frontiers in Psychology*, *15*, 1361616. https://doi.org/10.3389/fpsyg.2024.1361616

Liu, Q., Zhou, Y., et al. (2024). When ChatGPT is gone: Creativity reverts and homogeneity persists (No. arXiv:2401.06816). *arXiv*. https://doi.org/10.48550/arXiv.2401.06816

Luchini, S. A., Maliakkal, N. T., et al. (2025). Automated scoring of creative problem solving with large language models: A comparison of originality and quality ratings. *Psychology of Aesthetics, Creativity, and the Arts*. https://doi.org/10.1037/aca0000736

Moon, K., Green, A., et al. (2024). Homogenizing Effect of a Large Language Model (LLM) on Creative Diversity: An Empirical Comparison of Human and ChatGPT Writing. *arXiv*. https://doi.org/10.31234/osf.io/8p9wu

Organisciak, P., Acar, S., et al. (2023). Beyond semantic distance: Automated scoring of divergent thinking greatly improves with large language models. *Thinking Skills and Creativity*, *49*. https://doi.org/10.1016/j.tsc.2023.101356

Patterson, J., Pronchick, J., et al. (2024). *CAP: The Creativity Assessment Platform for online testing and automated scoring.*

Patterson, J. D., Barbot, B., et al. (2024). AuDrA: An automated drawing assessment platform for evaluating creativity. *Behavior Research Methods*, *56*(4), 3619-3636. https://doi.org/10.3758/s13428-023-02258-3

Pesout, O., & Nietfeld, J. L. (2021). How creative am I?: Examining judgments and predictors of creative performance. *Thinking Skills and Creativity*, *40*. https://doi.org/10.1016/j.tsc.2021.100836

Puente-Diaz, R. (2023). Metacognitive feelings as a source of information for the creative process: A conceptual exploration. *Journal of Intelligence*, *11*(3). https://doi.org/10.3390/jintelligence11030049

Puente-Díaz, R., Cavazos-Arroyo, J., et al. (2021). Metacognitive feelings as a source of information in the evaluation and selection of creative ideas. *Thinking Skills and Creativity*, *39*. https://doi.org/10.1016/j.tsc.2020.100767

Pustejovsky, J. E. T., E. (2018). Small sample methods for cluster-robust variance estimation and hypothesis testing in fixed effects models. *Journal of Business & Economic Statistics*, *36*(4), 672-683. https://doi.org/10.1080/07350015.2016.1247004

Rafner, J., Beaty, R. E., et al. (2023). Creativity in the age of generative AI. *Nature Human Behaviour*, *7*(11), 1836-1838. https://doi.org/10.1038/s41562-023-01751-1

Rietzschel, E. F., Ritter, S. M., et al. (2024). A systematic review of creativity evaluation and creativity selection measurement tasks. *Psychology of Aesthetics, Creativity, and the Arts*. https://doi.org/10.1037/aca0000638

Sidi, Y., Torgovitsky, I., et al. (2020). You may be more original than you think: Predictable biases in self-assessment of originality. *Acta Psychologica*, *203*, 103002. https://doi.org/10.1016/j.actpsy.2019.103002

Sijbom, R. B. L., Anseel, F., et al. (2017). Why seeking feedback from diverse sources may not be sufficient for stimulating creativity: The role of performance dynamism and creative time pressure. *Journal of Organizational Behavior*, *39*(3), 355-368. https://doi.org/10.1002/job.2235

Sowden, P. T., Pringle, A., et al. (2014). The shifting sands of creative thinking: Connections to dual-process theory. *Thinking & Reasoning*, *21*(1), 40-60. https://doi.org/10.1080/13546783.2014.885464

Torrance, E. P. (1972). Predictive validity of the Torrance tests of creativie thinking. *The Journal of Creative Behavior*, *6*, 236-252. https://doi.org/10.1002/j.2162-6057.1972.tb00936.x

Urban, K., & Urban, M. (2023). How can we measure metacognition in creative problem-solving? Standardization of the MCPS scale. *Thinking Skills and Creativity*, *49*. https://doi.org/10.1016/j.tsc.2023.101345

Urban, M., Děchtěrenko, F., et al. (2024). ChatGPT improves creative problem-solving performance in university students: An experimental study. *Computers & Education*, *215*. https://doi.org/10.1016/j.compedu.2024.105031

Urban, M., & Urban, K. (2021). Unskilled but aware of it? Cluster analysis of creative metacognition from preschool age to early adulthood. *Journal of Creative Behavior*, 1-9. https://doi.org/10.1002/jocb.499

Urban, M., & Urban, K. (2023). Do we need metacognition for creativity? A necessary condition analysis of creative metacognition. *Psychology of Aesthetics, Creativity, and the Arts*. https://doi.org/10.1037/aca0000647

Volle, E. (2018). Associative and Controlled Cognition in Divergent Thinking: Theoretical, Experimental, Neuroimaging Evidence, and New Directions. In *The Cambridge Handbook of the Neuroscience of Creativity* (pp. 333-360). https://doi.org/10.1017/9781316556238.020

Ward, T. B., Smith, S. M., et al. (1998). Creative Cognition. In *Handbook of Creativity* (pp. 189-212). https://doi.org/10.1017/cbo9780511807916.012

Wisniewski, B., Zierer, K., et al. (2019). The power of feedback revisited: A meta-analysis of educational feedback research. *Frontiers in Psychology*, *10*, 3087. https://doi.org/10.3389/fpsyg.2019.03087

Zhou, E., & Lee, D. (2024). Generative artificial intelligence, human creativity, and art. *PNAS Nexus*, *3*(3), pgae052. https://doi.org/10.1093/pnasnexus/pgae052

Supplementary Materials

Supporting text.

Fine-Tuned Large Language Model for Originality Scoring in Study 1 and 2.

Automated Detection of Invalid Responses in Study 1.

Tables S1 to S9

Supplementary Materials References

Supporting text

Fine-Tuned Large Language Model for Originality Scoring in Study 1 and Study 2. In Study 1, sixteen design problems across three categories—Social and Environmental Systems, Ability Difference and Limitation, and Transportation and Mobility—were rated for creativity by 20 expert raters, focusing on uncommonness, remoteness, and cleverness. Each rater assessed half of the problems using a planned missing data design. Responses were formatted as “An original solution to [problem] is [solution], rating: [score].” We selected three problems, one from each category, based on high factor loadings and strong correlations with human ratings. The GPT-3 davinci-002 model was fine-tuned using an 85/15 train/test split (n=6,549 training samples; n=1,156 test samples), across three epochs, with a batch size of 13 and a learning rate multiplier of 2. Model performance was evaluated via Pearson correlations between predicted and actual ratings on the held-out test set.

A large sample of first-year psychology students (*N* = 141, 85% female, age 18-28, *M* = 20.83, *SD* = 3.32) served as coders for randomly selected 200 drawings created in Study 2. All student-coders participated in a psychology of creativity course, where they were provided with information on creativity theories, assessment methods, and criteria for scoring creativity. The drawings were presented randomly in four sets (50 each), and coders used a 1-5 Likert scale (*1 = not original, 5 = highly original*). The coders' reliability was excellent (α = .995), so the ratings for each drawing were averaged. They ranged from a minimum of 1.22 to a maximum of 4.74, with *M* = 3.00 and *SD* = 0.97. Pearson correlation between the average coders' ratings and AI assessment of drawings originality was estimated at $r(200) = .924$, $p < .001$.

Automated Detection of Invalid Responses in Study 1. Given the substantial workload involved in manually screening large numbers of responses for irrelevant or nonsensical ideas, we decided to utilize modern techniques from natural language processing and deep learning to aid us in screening participant responses. This decision was based on recently developed methods utilizing bag-of-words based models to automatically screen for careless responding in social-behavioral data (Yeung & Fernandes, 2022). However, rather than relying on bag of words models that have limited ability to capture deeper semantic relationships, we instead utilized a contextual, transformer-based language model (TLM) that we finetuned to predict whether a given response was invalid, as we believed such a model would be more accurate in detecting careless responses.

Our model, based on the *deberta-base* architecture (He et al., 2021), was trained on a corpus of creative design problems rated as valid or invalid (NA) by five human raters. We considered a response invalid (1) if four or more raters marked it invalid and marked the response valid (0) otherwise. We split the ratings into train, dev, and two test sets. The first test set was an out-of-item (T_{oi}) set, constructed by holding out 15% of the items from the full dataset at random. The second test set was an out-of-response (T_{or}) set that was created by splitting the remaining data using an 80 / 10 / 10 (train / dev / T_{or}) ratio. T_{or} contained unseen responses to previously seen items during training, making it a weaker but easier test of generalization than T_{oi} (where both items and responses were unseen). The final dataset used to train the model contained 5,504 invalid and 5,196 valid responses in the training set, while the development set had 21 invalid and 640 valid responses. The two datasets were similarly imbalanced, with 19 invalid and 641 valid responses in the T_{or} set and 69 invalid and 1,736 valid responses in the T_{oi} set. We further expanded this dataset by applying a series of data augmentations that

transformed valid responses to become invalid. This consisted of introducing misspellings, wrong words, and cutting off responses, each of which likely rendered valid responses invalid. We applied each transformation to every valid response in the train set, leaving the other sets as is, and added the transformed valids to the invalid split. This brought the number of invalids to be roughly equal to the number of valids.

The model was trained using a standard binary cross entropy loss function, weighted proportionally to each class frequency. We used the dev set to tune the batch size, learning rate, number of training epochs, warmup steps, and the weight decay parameter. We implemented the model and training process using HuggingFace (Wolf et al., 2020) and evaluated the final model on both our test sets. Our final model weights and training scripts are available on our OSF repository. The model achieved a precision of 0.325 on the out-of-response (P_{or}) test set and 0.313 on the out-of-item (P_{oi}) test set. Recall was notably higher, with scores of 0.684 for recall of the out-of-response set (R_{or}) and 0.667 for recall of the out-of-item set (R_{oi}), indicating that while the model successfully identified a substantial portion of invalid responses, precision remained more moderate. These indicated that the model correctly identified a large fraction of true invalids without introducing too many false positive predictions. Based on this analysis, we concluded that the model would be suitable for performing a first pass rating of the data gathered in the main study but opted to include an additional round of human review for all responses marked NA.

Table S1.

Sociodemographic characteristics by engineering status (engineers, non-engineers) in Study 1

	Engineers		Non-engineers	
	M	*SD*	*M*	*SD*
Age	26.10	7.70	35.20	11.70
Years of experience in engineering	4.58	3.57		
Familiarity with design problems	3.38	1.01	2.97	1.06
Memory	3.07	1.40	3.20	1.52
Pedestrian	3.41	1.45	2.67	1.29
Water	3.64	1.38	3.03	1.38
	n	%	*n*	%
Total	327		227	
Identified sex / gender				
Female	88	27	119	52
Male	235	72	108	48
Other	0	0	0	0
No answer	4	1	0	0
Education				
Less than high school	1	<1	1	<1
High school graduate (or GED)	48	15	32	14
Some college, but no degree	70	21	35	15
Associate or bachelor's degree	133	41	118	52
Master's degree	71	22	35	15
Doctoral or professional degree	4	1	6	3
No answer	0	0	0	0
Engineering field				
Civil	36	11		
Electrical	61	19		
Industrial	47	14		
Mechanical	59	18		
Other	120	37		
None	4	1		

Note. Socio-demographic data are presented for participants used for analysis. Memory problem is "assist people with memory impairments remember important tasks". Pedestrian problem is "improve the safety of pedestrian crossings". Water problem is "improve access to clean water in remote areas". Average familiarity with design problems is higher for engineers ($M = 3.38$, $SD = 1.01$) than for non-engineers ($M = 2.97$, $SD = 1.06$), $t(552) = 4.59$, $p < .001$.

Table S2.

Correlations between originality and self-rating scores by condition (feedback, intermediate, control) and engineering status (engineers, non-engineers) in Study 1

	Engineering status		
Condition	Engineers	Non-engineers	Total
Feedback	.63***	.52***	.59***
Intermediate	.21*	.15*	.17*
Control	.20*	.26*	.22*

Notes. Scores are aggregated to the participant level. * $p < .05$. ** $p < .01$. *** $p < .001$.

Table S3.
Estimates of the predictors of originality scores for Linear Mixed Model 1, 2 and 3 in Study 1

	Model 1		Model 2		Model 3	
Fixed effects	b (SE)	*p*	b (SE)	*p*	b (SE)	*p*
Intercept	6.13 (0.15)	< .001	6.18 (0.17)	< .001	3.75 (0.23)	< .001
Condition (intermediate)	**0.67 (0.15)**	**< .001**	**1.06 (0.21)**	**< .001**	**0.92 (0.24)**	**< .001**
(feedback)	**0.54 (0.15)**	**< .001**	**0.66 (0.20)**	**< .001**	0.12 (0.24)	.62
Response order			0.08 (0.02)	.64	-0.02 (0.02)	.25
Response order × Condition (intermediate)			-0.03 (0.03)	.19	0.01 (0.02)	.68
(feedback)			-0.05 (0.03)	.06	0.02 (0.02)	.36
Trial order			-0.04 (0.04)	.31	-0.04 (0.03)	.28
Trial order × Condition (intermediate)			-0.14 (0.06)	.01	-0.09 (0.05)	.06
(feedback)			0.01 (0.06)	.86	0.08 (0.05)	.11
Self-ratings					**0.41 (0.02)**	**< .001**
Self-ratings × Condition (intermediate)					-0.01 (0.02)	.69
(feedback)					**0.11 (0.03)**	**< .001**
Random effects	σ^2	*SD*	σ^2	*SD*	σ^2	*SD*
Participant (intercept)	1.97	1.40	1.95	1.40	1.76	1.33
Problem (intercept)	0.03	0.17	0.03	0.17	0.08	0.28
Model comparison				*p*		*p*
AIC	28,104		28,087		26,483	
$\Delta\chi^2$ (*df*)			29.07 (6)	< .001	1610.29 (3)	< .001

Model 1. Originality ~ condition + (1 | participant) + (1 | problem)
Model 2. Originality ~ condition + response order + trial order + condition:response order + condition:trial order + (1 | participant) + (1 | problem)
Model 3. Originality ~ condition + response order + trial order + self-ratings + condition:response order + condition:trial order + condition:self-ratings + (1 | participant) + (1 | problem)
Notes. Significant fixed effects are in bold. Condition is a three-level categorical variable with "control" as the reference category. Results for fixed effects are presented as regression coefficients (b) with standard errors (SE) in parentheses. χ^2 indicates chi-square statistic of the likelihood ratio test. σ^2 indicates the variance estimate of the random intercept. AIC indicates Akaike's Information Criterion.

Table S4.
Estimates of the multilevel mediation model in Study 1

Effect		b	95% CI	*p*
Within				
	Condition - Self-rating	1.49	[-0.10, 3.09]	.07
	Condition - Originality	1.16	[-0.23, 2.55]	.10
	Self-rating - Originality	**0.42**	**[0.40, 0.45]**	**< .001**
	Indirect (Condition - Self-rating - Originality)	0.63	[-0.04, 1.31]	.07
Between				
	Condition - Self-rating	**-0.25**	**[-0.38, -0.12]**	**< .001**
	Condition - Originality	**0.34**	**[0.20, 0.49]**	**< .001**
	Self-rating - Originality	**0.38**	**[0.27, 0.48]**	**< .001**
	Indirect (Condition – Self-rating- Originality)	**-0.10**	**[-0.15, -0.04]**	**< .001**

Notes. Significant fixed effects are in bold. Results for within and between effects are presented as regression coefficients (b)

Table S5.
Correlations between originality and self-rating scores by condition (feedback, control) and response order (first, second drawing) in Study 2

	Response order		
Condition	First drawing	Second drawing	Total
Feedback	.15***	.22***	.16***
Control	.15***	.15***	.12***

Notes. Scores are aggregated to the participant level. *** $p < .001$. In the "feedback" condition, participants received feedback after the first drawing, while in the "control" condition, feedback was only provided after the second drawing, meaning the second idea was not influenced by feedback. In both conditions, the first drawing serves as the baseline, with the key comparison focusing on the second drawing in each condition.

Table S6.
Estimates of the predictors of originality scores for Linear Mixed Model 1, 2 and 3 in Study 2

	Model 1		Model 2		Model 3	
Fixed effects	b (SE)	*p*	b (SE)	*p*	b (SE)	*p*
Intercept	4.49 (0.16)	< .001	4.23 (0.16)	< .001	3.34 (0.17)	< .001
Condition	**0.16 (0.02)**	**< .001**	**0.10 (0.04)**	**.01**	**0.12 (0.06)**	**.04**
Response order			**-0.18 (0.02)**	**< .001**	**-0.17 (0.05)**	**< .001**
Response order × Condition			**0.41 (0.03)**	**< .001**	**0.18 (0.08)**	**.02**
Trial order			**0.28 (0.02)**	**< .001**	**0.30 (0.02)**	**< .001**
Trial order × Condition			**-0.06 (0.02)**	**<.01**	**-0.06 (0.02)**	**.01**
Self-ratings					**0.12 (0.01)**	**< .001**
Self-ratings × Condition					-0.01 (0.01)	.50
Self-ratings × Condition × Response order					**0.04 (0.01)**	**< .001**
Random effects	σ^2	*SD*	σ^2	*SD*	σ^2	*SD*
Participant (intercept)	2.03	1.43	2.08	1.45	2.06	1.44
Shape (intercept)	0.16	0.40	0.16	0.40	0.16	0.40
Model comparison				*p*		*p*
AIC	340,867		340,256		338,183	
$\Delta\chi^2$ (*df*)			618.55 (4)	< .001	2081.51(4)	< .001

Model 1. Originality ~ condition + (1 | participant) + (1 | shape)
Model 2. Originality ~ condition + response order + trial order + condition:response order + condition:trial order + (1 | participant) + (1 | shape)
Model 3. Originality ~ condition + response order + trial order + self-ratings + condition:response order + condition:trial order + condition:self-ratings + (1 | participant) + (1 | shape)
Notes. Significant fixed effects are in bold. Condition is a two-level categorical variable with "control" as the reference category. Response order is a two-level categorical variable with "first drawing" as the reference category. Results for fixed effects are presented as regression coefficients (b) with standard errors (SE) in parentheses. χ^2 indicates chi-square statistic of the likelihood ratio test. σ^2 indicates the variance estimate of the random intercept. AIC indicates Akaike's Information Criterion.

Table S7.
Estimates of the multilevel mediation model in Study 2

Effect	b	95% CI	*p*
Self-rating [2nd] ~ Condition (a path)	**-0.14**	**[-0.22, -0.06]**	**< .001**
(Self-rating [2nd] ~ Self-rating [1st])	**0.43**	**[0.42, 0.45]**	**< .001**
Originality [2nd] ~ Condition (c path)	**0.54**	**[0.48, 0.60]**	**< .001**
Originality [2nd] ~ Self-rating [2nd] (b path)	**0.14**	**[0.13, 0.15]**	**< .001**
(Originality [2nd] ~ Originality [1st])	**0.48**	**[0.47, 0.50]**	**< .001**
Indirect effect (Condition - Self-rating [2nd] - Originality [2nd])	**-0.02**	**[-0.03, -0.01]**	**< .001**

Notes. Significant fixed effects are in bold. [1st] and [2nd] indicate the first and second drawings, respectively. Results are presented as regression coefficients (b).

Table S8.
(Preregistered analysis) Estimates of the predictors of accuracy scores for Generalized Linear Mixed Model 1 and 2 in Study 1

	Model 1		Model 2	
Fixed effects	b (SE)	*p*	b (SE)	*p*
Intercept	0.22 (0.01)	< .001	0.22 (0.01)	< .001
Condition (intermediate)	0.01 (0.01)	.52	0.03 (0.02)	.06
(feedback)	**0.04 (0.01)**	**< .001**	**0.05 (0.02)**	**< .01**
Response order			0.00 (0.01)	.53
Response order × Condition (intermediate)			0.00 (0.01)	.68
(feedback)			0.00 (0.01)	.30
Trial order			0.00 (0.01)	.56
Trial order × Condition (intermediate)			**-0.01 (0.01)**	**< .01**
(feedback)			0.00 (0.01)	.92
Random effects	σ^2	*SD*	σ^2	*SD*
Participant (intercept)	0.01	0.11	0.01	0.12
Problem (intercept)	0.00	0.01	0.00	0.01
Model comparison				*p*
AIC	37,024		37,019	
$\Delta\chi^2$ (*df*)			17.22 (6)	< .001

Model 1. Accuracy ~ condition + (1 | participant) + (1 | problem)
Model 2. Accuracy ~ condition + response order + trial order + condition:response order + condition:trial order + (1 | participant) + (1 | problem)
Notes. GLMM generated using a gamma distribution. Significant fixed effects are in bold. Condition is a three-level categorical variable with "control" as the reference category. Results for fixed effects are presented as regression coefficients (b) with standard errors (SE) in parentheses. X^2 indicates chi-square statistic of the likelihood ratio test. σ^2 indicates the variance estimate of the random intercept. AIC indicates Akaike's Information Criterion.

Table S9.
(Preregistered analysis) Estimates of the predictors of accuracy scores for Generalized Linear Mixed Model 1 and 2 in Study 2

	Model 1		Model 2	
Fixed effects	b (SE)	*p*	b (SE)	*p*
Intercept	0.05 (0.00)	< .001	0.05 (0.00)	< .001
Condition	**0.002 (0.00)**	**< .001**	-0.003 (0.00)	.09
Response order			**-0.001 (0.00)**	**<.01**
Response order × Condition			**0.003 (0.00)**	**< .001**
Trial order			**0.004 (0.00)**	**< .01**
Trial order × Condition			0.00 (0.00)	.50
Random effects	σ^2	*SD*	σ^2	*SD*
Participant (intercept)	0.00	0.00	0.00	0.00
Shape (intercept)	0.00	0.00	0.00	0.00
Model comparison				*p*
AIC	579,450		579,283	
$\Delta\chi^2$ (*df*)			175.55 (4)	< .001

Model 1. Accuracy ~ condition + (1 | participant) + (1 | shape)
Model 2. Accuracy ~ condition + response order + trial order + condition:response order + condition:trial order + (1 | participant) + (1 | shape)
Notes. GLMM generated using a gamma distribution. Significant fixed effects are in bold. Condition is a two-level categorical variable with "control" as the reference category. Response order is a two-level categorical variable with "first drawing" as the reference category. Results for fixed effects are presented as regression coefficients (b) with standard errors (SE) in parentheses. χ^2 indicates chi-square statistic of the likelihood ratio test. σ^2 indicates the variance estimate of the random intercept. AIC indicates Akaike's Information Criterion.

Supplementary materials References

He, P., Liu, X., Gao, J., & Chen, W. (2021). DeBERTa: Decoding-enhanced BERT with Disentangled Attention. https://doi.org/10.48550/arXiv.2006.03654

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., Davison, J., Shleifer, S., von Platen, P., Ma, C., Jernite, Y., Plu, J., Xu, C., Le Scao, T., Gugger, S., & Rush, A. (2020). Transformers: State-of-the-Art Natural Language Processing. Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations,

Yeung, R. C., & Fernandes, M. A. (2022). Machine learning to detect invalid text responses: Validation and comparison to existing detection methods. *Behavior Research Methods*, *54*(6), 3055-3070. https://doi.org/10.3758/s13428-022-01801-y