WORKING PAPER · V4 · 6 SEPTEMBER 2026 · DEPOSITED ON ZENODO

Stimulate, Don't Replace: How AI Design Shapes Human Creative Effort

The CineCoach Case and Recommendations for European Policy

Ajay Bakshi[1] (ORCID 0009-0009-2448-4031), Véronique Briquet-Laugier[1] (ORCID 0009-0000-6011-2416), Paco Wiser[1,2] (ORCID 0009-0005-6672-4989)

[1] DECCS, Paris · [2] EICAR, Paris · Ajay Bakshi MBBS, M.Ch.(Neurosurgery) · Véronique Briquet-Laugier PhD · Correspondence: info@cinecoach.paris · **Declaration of interest:** the authors are the founders of CineCoach, the system described in Section 5.

ABSTRACT

Films from the 27 Member States of the European Union, counted with their predecessor states, have won 52 of the 79 Academy Awards for Best International Feature Film and are credited on 54 of the 102 winners of the Palme d'Or. The festivals of Venice, Cannes and Berlin are European. Some 8.9 million Europeans work in culture in 24 official languages. Generative AI now touches every one of these fields.

The European Commission is preparing an AI strategy for the cultural and creative sectors on the principle that AI should complement and not replace human creativity. We ask what the evidence says about how that principle can be built into a tool and what it implies for public policy. We review seven bodies of research (neural and cognitive models of creativity, the generation effect, motivation and evaluation, deliberate practice and feedback, tutoring and Socratic instruction, guided discovery and human–AI co-creation) and report only findings from sources we opened with their statistics as stated.

Human–AI co-creation raises individual output and lowers collective diversity (meta-analytic $g = 0.27$ and $g = -0.86$ with a second meta-analysis putting homogenisation at $d = 0.33$). The interface decides which effect dominates – an AI that questions the person about their own idea preserves diversity and ownership while an AI that rewrites the idea loses both ($d = 0.76$ and 0.57). In learning an answer-giving AI raises assisted performance and lowers unassisted performance (–17%) while a hint-giving AI does no harm. High-information feedback has twice the effect of corrective feedback ($d = 0.99$ against 0.46).

We also report the findings that cut against a simple Socratic thesis – a randomised trial in which a strategy-giving AI coach left users worse off than no AI at all, another with no difference between

Socratic and non-Socratic tutors and the absence of any reliable neural signature for divergent thinking.

Used wrongly AI will flatten the plurality that makes European creativity unique. Used rightly it will augment and accelerate European creativity. We describe CineCoach (a coaching system encoded from one cinematographer's method and running in fifteen languages, eleven of them official EU languages, each with its own natively analysed film curation) as one working example of the augmenting design and we state its limits. We close with an agenda for the strategy – fund the design rather than the technology, measure output diversity as an outcome and put artist-facing AI under the control of the schools and studios that use it.

Keywords: generative AI, creativity, homogenisation, film education, Socratic tutoring, feedback, cultural diversity, European cultural policy

1. Introduction

Public argument about artificial intelligence and the arts has settled on a question that cannot be answered and does not need to be – whether a machine can be creative. The European Commission's call for evidence for an AI strategy for the cultural and creative sectors (published 27 July 2026 with feedback open to 11 September) sets that question aside and states a design goal instead – "AI should complement and not replace human creativity and agency" (European Commission 2026). Ministers meeting informally in Nicosia on 2 June 2026 asked how public policy can ensure that AI "functions as a supportive tool without undermining artistic freedom, authorship, transparency and the fair recognition and remuneration of creators and cultural professionals".

What is at stake in that goal is concrete. Films from European countries, Europe being defined throughout this paper as the 27 Member States of the European Union together with their predecessor states (European Union, n.d., Countries; the counting rule and its alternatives are stated in the reference notes), have won 52 of the 79 awards recorded in the by-country tally for Best International Feature Film and 213 of its 359 nominations (Italy 14 wins and France 12). A Member State is credited as a production country on 54 of the 102 films that have won the Palme d'Or. The festivals of Venice (founded 1932), Cannes (first held 1946, after the edition decreed for 1939 was cancelled by the war) and Berlin (1951) are European. In 2025 some 8.9 million people worked in culture in the European Union – 4.3% of all employment and among them 1.8 million artists and writers (Eurostat 2026). The cultural and creative industries account for about 3.95% of EU value added (European Commission DG GROW, undated). The Union works in 24 official languages. Among the world's hundred largest luxury-goods companies, Italy has 23 and seven

French houses take nearly a third of the sales (Deloitte 2023). One design of generative AI erodes that plurality and another compounds it (Section 4).

In the Commission's 2025 survey (Special Eurobarometer 562 with fieldwork in February and March 2025) 81% preferred human-made cultural content to AI-generated content (the figure in the report's narrative; its annex table gives 77%) and 73% were concerned about the effect of generative AI on artists' employment or earnings.

Under what conditions does an AI system augment an artist's own creative work and under what conditions does it substitute for it? If the answer depends on how the system is built rather than on the underlying model then "complement, not replace" becomes a specification – and a specification can be funded, procured, taught and measured.

This paper reviews the evidence for that claim. We write as practitioners with a declared interest (we build CineCoach, an AI coach for screenwriting and cinematography students which we use in Section 5 as a case). The review in Sections 3 and 4 stands independently of the product. We have reported the findings that cut against our own thesis wherever the scans found them. Section 6 draws the policy consequences for Europe.

One of us (Bakshi) trained as a neurosurgeon and spent years in neuroscience research. The brain science in this paper stops at Section 3.1. A review of 72 brain-imaging experiments of creativity found that the neuroimaging studies of divergent thinking among them, 14 in all, show "no reliable changes above and beyond diffuse prefrontal activation" (Dietrich and Kanso 2010). Adding logically irrelevant neuroscience information to a weak explanation makes non-expert readers rate it as more satisfying (Weisberg et al. 2008). Later work shows the pull is not specific to brain language but reflects a general preference for reductive detail across the sciences (Hopkins, Weisberg and Taylor 2016). So Section 3.1 takes two things from the cognitive literature – the mechanism (generation and evaluation as separate processes, with new ideas assembled from what a person already knows) and the vocabulary. The policy argument rests on the behavioural studies.

2. Method

This is a critical integrative review rather than a systematic one. We did not register a protocol or fix a database list or use dual screening. We describe what we did so that a reader can judge the coverage.

On 28 August 2026 we ran seven structured literature scans (one for each body of work in Table 1). Each scan used web search to locate meta-analyses and reviews first and then landmark empirical

studies and then recent studies of human–AI interaction (2023 to 2026). For every source the scan opened the abstract or the indexed record or the full text and recorded statistics only as stated in the opened text. Sources that could not be opened are listed as such and contribute no numbers. Each scan was instructed to look for findings that contradict our thesis and to report them. The scans were run by a large language model under our direction. We set the questions, the search order and the rule that no number may be reported unless it appears in the opened text. A second pass on 28 August 2026 re-checked 76 statistics, quotations and attributed findings against their sources. Its corrections are in this text and are carried into the deposited evidence map.

We then described our own system (Section 5) from two records: the system as it is built, and the transcripts and review sessions from which we encoded the expert's method. A feature inventory with source references forms section 8 of the deposited evidence map.

BODY OF WORK	QUESTION ASKED	ENTRIES	OF WHICH CUT AGAINST THESIS
Neural and cognitive models of creativity	What is known about generation and evaluation as processes?	14	4
Generation effect, desirable difficulties, AI offloading	Does producing beat receiving, and does AI change this?	12	4
Intrinsic motivation and evaluation	Does evaluation harm creativity, and which kind?	16	5
Deliberate practice and feedback	What kind of feedback builds skill?	11	6
Tutoring, self-explanation, Socratic AI	Does questioning work, and does it scale?	10	6
Guided discovery and expertise reversal	When should a coach tell, and when ask?	11	6
Human–AI co-creation and homogenisation	What does generative AI do to creative diversity?	13	4

Table 1. The seven scans. Entry counts include sources marked as not opened. The full evidence map, with each source's verification status and statistics, is deposited with this paper.

Several sources we rely on are preprints at the time of writing and are marked as such in the references: Holzner, Maier and Feuerriegel (2025), Maier, Schneider and Feuerriegel (2025), Kumar et al. (2024), de Rooij and Biskjaer (2026), Atari et al. (2023), Blasco and Charisi (2025), Kao, Grant and Woltering (2025) and Kosmyna et al. (2025).

3. The evidence

3.1 How ideas are generated

Cognitive models of creativity describe an alternation between generating candidates and evaluating them (Sowden, Pringle and Gabora 2015). Neuroimaging reviews associate creative cognition with cooperation between the brain's default network, linked to internally directed thought, and its executive control network (Beaty et al. 2016), and a coordinate-based meta-analysis of divergent-thinking studies locates the consistent activations in a semantic control network "important for flexible retrieval of stored knowledge" (Cogdell-Brooke et al. 2020). Generating a new idea, as distinct from recalling an old one, recruits left inferior parietal cortex, a region tied to mental simulation and future thought. Idea generation as a whole, new and recalled alike, looks like internally directed attention with controlled semantic retrieval (Benedek et al. 2014). The raw material of a person's creative output is what that person already knows and has lived (a grandmother's kitchen, a first job, a film seen at fourteen), recombined under attention.

Two further findings from memory research bear on coaching. A short induction that trains people to recall a recent experience in detail raised their subsequent divergent-thinking output ($d = 0.52$ and 0.77 in two samples of 23; Madore, Addis and Schacter 2015). And incubation, a pause between preparation and return, improves creative problem solving (weighted mean effect 0.29, 95% CI 0.21 to 0.39, across 117 studies), more so after longer preparation for visual and creative problems though not for linguistic ones (Sio and Ormerod 2009). A coach that sends the student back into their own experience, and that allows a pause before the next round, is working with these mechanisms rather than against them.

The network-coupling results come from trait comparisons in small or correlational samples ($n = 24$ in Beaty et al. 2014; external-sample correlations of 0.13 to 0.35 in Beaty et al. 2018). They say nothing about whether a teaching method changes the coupling. And the largest quantitative review of creativity training finds that the best programmes teach named heuristics with realistic practice. Lecture-based delivery correlated positively with effect size ($r = .20$) and discussion-based delivery did not ($r = -.04$), while courses stressing "unconstrained exploration" did worse (Scott, Leritz and Mumford 2004). A questioning coach must also teach method.

3.2 The generation effect, and what AI does to it

The generation effect (Slamecka and Graf named it in 1978) is one of the most replicated results in memory research – material a person produces is retained better than the same material read (Slamecka and Graf 1978; meta-analytic effect .40 across 86 studies, Bertsch et al. 2007; reliable across 310 experiments, McCurdy et al. 2020). Retrieval practice shows the same pattern – across 159 testing-versus-restudy comparisons the benefit is $g = 0.50$, and it is larger when the practice

test requires recall than when it requires recognition (Rowland 2014). A larger meta-analysis confirms the benefit over restudy (g = 0.51) and over no activity (g = 0.93), though it finds multiple-choice practice tests at least as effective as recall-based ones (Adesope, Trevisan and Sundararajan 2017). Bjork (1994) generalised these into "desirable difficulties": conditions that make learning harder in the moment, including reduced feedback frequency, improve later retention and transfer.

Generative AI is, by construction, a machine for removing this difficulty – the answer arrives before the effort. Recent studies show the predicted cost in classrooms whenever the outcome is measured without the tool. In a field experiment with nearly 1,000 secondary-school students, an unrestricted GPT tutor raised practice scores by 48% and lowered scores on a later unassisted exam by 17%. A tutor restricted to giving hints raised practice scores by 127% and left exam scores no different from controls (Bastani et al. 2025). University students writing with ChatGPT improved their essay scores but gained no more knowledge, and transferred it no better, than students supported by a human expert, by writing analytics tools, or by nothing at all – a pattern the authors interpret as "metacognitive laziness" (Fan et al. 2025). In a small study of essay writers, those using an LLM reported the lowest ownership of their essays and had difficulty quoting their own work (Kosmyna et al. 2025; a preprint with 54 participants, whose EEG findings we do not use, to be read as suggestive). A cross-sectional survey of 666 people found frequent AI use negatively correlated with critical thinking, mediated by cognitive offloading (Gerlich 2025).

Two recent meta-analyses report the opposite sign, and a reader should know why both can be true. Across 35 experimental studies with 4,193 participants, ChatGPT raised student learning outcomes by g = 0.670 (Wu et al. 2026), and across 37 studies it raised academic achievement by g = 0.577 (Liu, Zuo and Lu 2025). Neither synthesis separates performance measured with the tool from performance measured afterwards without it. Wu et al.'s inclusion criteria require only a control group or a pre-test and post-test, and neither synthesis lists assessment timing among its moderators. That separation is the one Bastani et al. made, and it reversed the sign. A third meta-analysis reporting a larger effect (Wang and Fan 2025) was retracted in April 2026 and we do not rely on it.

The evidence also shows that withholding output is only half the answer – the coach still has to teach something. An AI tutor built to give step-by-step explanations, sequenced through problems that students had to attempt, outperformed an active-learning physics class (effect size 0.63 on an immediate post-test; Kestin et al. 2025). And the hint-giving tutor in Bastani et al. reached parity with no-AI controls, not superiority. Producing must be preserved, and the coach still has to teach.

3.3 Evaluation and motivation

Amabile's programme showed that expecting evaluation lowered the judged creativity of collages made by students (Amabile 1979, as reported in Amabile 2012), that creative writers primed with extrinsic reasons wrote poems judged less creative than those primed with intrinsic reasons (means 15.74 vs 19.88 on the judges' scale, $F(2,69) = 4.66$; Amabile 1985), and that contracting for a reward lowered the creativity of children's stories and adults' collages (Amabile, Hennessey and Grossman 1986). Intrinsic motivation correlates with the creativity of produced work at $r = 0.30$ across 26 samples (de Jesus et al. 2013).

But the meta-analytic picture is more precise than "evaluation harms". Byron, Khazanchi and Nazarian (2010), across 76 experimental studies, found a curved relation – a low evaluative context raised creative performance above control, while a highly evaluative context lowered it. Creativity-contingent rewards raised creative performance, most when combined with positive, task-focused feedback and choice (Byron and Khazanchi 2012). Positive verbal feedback enhanced intrinsic motivation ($d = 0.33$), whereas tangible expected rewards undermined it (Deci, Koestner and Ryan 1999). The same limits set in a controlling style reduced children's intrinsic motivation and painting creativity, and set in an informational style they did not (Koestner et al. 1984). And written comments alone produced higher interest and performance on a divergent task than grades, or grades with comments (Butler 1988).

The reading for a coaching tool is "evaluate informationally, at the level of the task, without a grade". Mild evaluative attention (a reader who cares whether the scene works) is part of what makes the work matter.

3.4 Feedback

Feedback is among the strongest influences on learning – and its sign depends on its type. The classic meta-analysis found an average effect of $d = .41$ across 607 effect sizes, and that over a third of feedback interventions reduced performance, the harm concentrated where feedback shifted attention from the task to the self (Kluger and DeNisi 1996). The most recent large synthesis (435 studies, $k = 994$, $N > 61{,}000$) gives an overall $d = 0.48$, with reinforcement or punishment at 0.24, corrective feedback at 0.46, and high-information feedback at 0.99 (Wisniewski, Zierer and Hattie 2020). Hattie and Timperley (2007) had earlier reported cues at 1.10 and praise at 0.14.

Two small studies from film and design education add the practitioner's view. In a small action-research study with five foundation-year film and television production students, Thompson (2018) identified "critiquing without knowing what action to take" as a barrier to students using their feedback, and reported that feedback needed to be "straight to the point and clear" before students could build an action plan. Teachers on a vocational filmmaking course assessed work on

four principles, and their aesthetic judgement pulled against a syllabus focused on technical skill (Florén, Bezemer and Domingo 2025). A genre analysis of design-studio critique found that the feedback socialised students into egalitarian relationships and autonomous decision-making identities, which the authors judged more reflective of academic stages or idealised workplaces than of actual professional settings, and which they argue complicates the pre-professional purpose of the critique (Dannels and Martin 2008). Expert feedback in the arts is specific, actionable, and partly tacit. A tool that supplies the first two properties supplies a real and scarce good, and does not replace the third.

On expertise itself, the evidence is contested. Deliberate practice explained 24% of performance variance in games, 23% in music and 5% in education across one meta-analysis, and 14% across all domains (Macnamara, Hambrick and Oswald 2014, as corrected in 2018). Ericsson and Harwell (2019) re-analysed the same effect sizes under stricter criteria, keeping only studies with objectively reproducible performance and with solitary practice, whether guided by a teacher or self-directed. On the surviving effect sizes practice accounted for about 29% of the variance, or 61% once they corrected for the assumed unreliability of the practice and performance measures. That restricted set retained almost nothing outside games, music and sport – no studies from the professions and one from education. That original definition, practice designed by a teacher and guided by feedback, is what a coaching tool attempts to supply at scale.

3.5 Tutoring and Socratic questioning

Bloom (1984) reported that tutored students scored about two standard deviations above conventional classes, and called tutoring "too costly for most societies to bear on a large scale". Later synthesis reduced the estimate – human tutoring d = 0.79, step-based intelligent tutoring systems d = 0.76, answer-based systems d = 0.3 (VanLehn 2011), and a median of 0.66 across 50 evaluations of intelligent tutoring, shrinking against active controls (Kulik and Fletcher 2016). Prompting students to explain material to themselves has a medium effect (g = .55 across 69 effect sizes; Bisra et al. 2018), though a major review rated elaborative interrogation and self-explanation only of "moderate utility", noting they "have not been adequately evaluated in educational contexts" (Dunlosky et al. 2013). Instruction that combines authentic or situated problems, dialogue and mentoring produced the largest effect of the four combinations examined for critical thinking ($g+$ = 0.57, k = 19, 95% CI 0.38 to 0.77), against 0.23 for dialogue alone and 0.25 for authentic instruction alone (Abrami et al. 2015, Table 6).

Direct tests of Socratic AI are few, recent and mixed. A randomised comparison of Socratic and non-Socratic AI tutors, with 122 students aged 14 to 16 in Brussels and Seville, "found no

significant difference in student performance", and reactions to the Socratic tutor divided rather than settled – 44% called it helpful or very helpful against 21% who called it not at all helpful (Blasco and Charisi 2025). The same trial ran a second randomised manipulation, and there step-by-step explanations did improve performance. A preprint reports gains in argumentation and critical thinking for an inquiry-plus-AI arm among 90 tenth-graders (Kao, Grant and Woltering 2025). No source we opened tests tutoring, self-explanation or Socratic AI on a creative-production task such as screenwriting.

3.6 Guidance and the expertise reversal effect

Minimal guidance is less effective than strong guidance for learners without prior knowledge (the finding that started the "discovery learning" argument in 2006), and "the advantage of guidance begins to recede only when learners have sufficiently high prior knowledge" (Kirschner, Sweller and Clark 2006). Across two meta-analyses drawing on 164 studies, unassisted discovery lost to explicit instruction ($d = -0.38$ across 108 studies) while enhanced discovery, supported by feedback, scaffolding, worked examples and elicited explanations, beat other forms of instruction ($d = 0.30$ across 56 studies; Alfieri et al. 2011). Guidance in inquiry learning improves outcomes ($d = 0.50$; Lazonder and Harmsen 2016). Techniques that help novices can harm experts, the expertise reversal effect (Kalyuga et al. 2003), and the reversal depends on the complexity of the task relative to the learner's knowledge rather than on a fixed label (Chen, Kalyuga and Sweller 2017).

The evidence resists a simple "tell beginners, ask advanced" rule. Problem solving before instruction beats instruction before problem solving for learning new concepts ($g = 0.36$ across 53 studies), though the effect reverses for young children and for domain-general skills (Sinha and Kapur 2021). Adding self-explanation prompts to worked examples in mathematics reduced their benefit (Barbieri et al. 2023). The defensible statement is that the amount of scaffolding should track the learner's prior knowledge and the task's complexity, and that a coaching system should adapt per dimension of skill rather than per level of the student.

4. Two designs for the same model

The human–AI co-creation literature now supports one conclusion with meta-analytic weight and one with experimental weight.

First, generative AI raises individual creative output and lowers collective creative diversity. Across 28 studies and 8,214 participants, humans working with generative AI outperformed humans alone on creative performance ($g = 0.27$), while idea diversity fell sharply ($g = -0.86$), and AI alone

did not differ from humans alone (Holzner, Maier and Feuerriegel 2025). A three-level meta-analysis of 19 studies and 61 effect sizes finds a homogenisation effect of d = 0.33, robust to sensitivity analysis and not explained by publication bias in the 14 peer-reviewed studies tested, which the authors describe as "a gradual, scale-dependent reorganization of creative diversity", "subtle in isolated instances, but consequential when examined at scale and over time" (de Rooij and Biskjaer 2026). The single-study evidence agrees – writers given AI story ideas produced stories rated more creative, most for less creative writers, and more similar to one another (Doshi and Hauser 2024). ChatGPT users produced more ideas that were less distinct across users and felt less responsible for them (Anderson, Shah and Kreminski 2024). Across more than four million artworks by over 50,000 users of one art-sharing platform, adopters of image generators raised their output by 25% and the value of their work by 50% measured as favourites per view, while the average novelty of their artworks fell relative to non-adopters and their subjects and styles grew steadily more alike (Zhou and Lee 2024). The effect is cultural – writing suggestions from an AI assistant pushed Indian writers toward Western styles and diminished the cultural nuances in their writing (Agarwal, Naaman and Vashistha 2025), and large language models' responses resemble those of Western, educated, industrialised populations, with similarity falling as cultural distance grows (r = –.70; Atari et al. 2023).

Second, the interface decides which of these effects dominates. Maier, Schneider and Feuerriegel (2025) compared, in a pre-registered experiment with 486 participants, an LLM that asked the person questions about their own idea (human-led) with one that rewrote the idea for them (model-led). Question mode raised idea quality against a no-AI control (d = 0.36) and left idea diversity no lower than a plain-LLM baseline (d = 0.003, not significant), while giving users higher perceived ownership than either the rewrite mode (d = 0.57) or a plain LLM (d = 0.60). Rewrite mode also raised quality against the control (d = 0.55), but produced a less diverse idea pool than question mode (d = 0.76). In a coded subset of 20 conversations per condition, 90% of rewrite-mode participants accepted the model's rewrite without further elaboration, while in question mode 60% substantially elaborated their own idea and a further 40% added a new one.

The two conditions used the same model. The only difference was whether the system produced the artefact or interrogated its author.

Four findings qualify this. Kumar et al. (2024), in randomised experiments with 1,100 participants across two studies, found that an LLM that gave strategies rather than answers, a "coach-like" design, left users worse off on later unassisted divergent thinking than users with no AI exposure at all (p = 0.009, in the study of 460 participants that tested divergent thinking), and reduced idea diversity both during and after use. The comparison with answer-giving was not significant (p =

0.126). Their coach supplied lists of creative strategies, and Maier et al.'s asked questions about the person's own idea. The two coaches differ in kind, and the two studies differ in what they measured: Kumar et al. tested a delayed unaided outcome and Maier et al. did not. No study has yet tested questions anchored in the person's own material on a delayed unaided outcome. We therefore treat the difference between generic guidance and anchored questions as a hypothesis, not a result. Section 5 treats the wording of a coach's prompts as a design decision for that reason, and Recommendation 5 asks for the study that would test it. Hou et al. (2025) found that output-producing AI raised novelty by 76% at the ideation stage for everyone, while giving experts nothing at implementation and costing them 57% more time – augmentation is stage-dependent. And the pooled diversity loss in Holzner et al. cannot be attributed to interface type, because the meta-analysis does not separate producing from questioning designs. Expertise still sets the ceiling – images from professional artists' prompts were rated more creative than those from a chatbot's, which beat novices' (Seli et al. 2025), and a structured co-creation process widened the gap between experienced and novice designers on novelty and quality (Wang et al. 2025).

Table 2 states the distinction as a specification.

PROPERTY	GENERATING DESIGN	QUESTIONING DESIGN	EVIDENCE	TASK AND OUTCOME MEASURED
Who produces the artefact	The system	The person	Maier 2025; Bastani 2025	Ideation, immediate (Maier); school mathematics, delayed exam (Bastani)
Effect on individual quality	Larger gain (d = 0.55)	Smaller gain (d = 0.36)	Maier 2025	Ideation, immediate
Effect on diversity across people	Falls (g = −0.86 pooled; d = 0.76 vs questioning)	Preserved at the plain-LLM level (d = 0.003)	Holzner 2025; Maier 2025	Pooled across tasks (Holzner); ideation, immediate (Maier)
Ownership	Falls (d = 0.57)	Preserved	Maier 2025; Anderson 2024	Ideation, immediate
Later unassisted skill	Falls (−17%)	Unchanged with hints; falls with generic strategy lists; untested for anchored questions	Bastani 2025; Kumar 2024	School mathematics, delayed exam (Bastani); divergent thinking, delayed unaided (Kumar)
Anchor of the intervention	The model's distribution	The person's own material and experience	Kumar 2024 vs Maier 2025; Madore 2015	Divergent thinking, delayed (Kumar); ideation, immediate (Maier); alternate uses, immediate (Madore)

Feedback type	A finished product	High-information, task-level	Wisniewski 2020; Kluger and DeNisi 1996	Educational feedback, pooled across school and university tasks

Table 2. The design distinction. Effect sizes as stated in the opened sources. No row rests on a creative-production task with a delayed unaided outcome (Section 5.4). See the evidence map for verification status.

5. CineCoach

Our own system is an instance of the questioning design. It is a description of what we built and not a test of whether it works.

5.1 Capturing the method

Paco Wiser is a cinematographer trained at INSAS in Brussels and at the Sorbonne under Pierre Jenn, with 45 years in cinema and a faculty position at EICAR, Paris. The problem was to move his judgement into a machine without flattening it into general film-school advice, which is the homogenising failure of Section 4 stated as a design risk. We followed a protocol of our own in three stages.

Extract. We recorded four sessions with him between 19 February and 7 March 2026, about three hours in total, and transcribed them in full. The sessions did not open by asking for principles. They asked him to read particular scenes and say what was wrong with them, and only then to explain the values behind the diagnosis. We were looking for five things: what he judges true or false in a scene, the order in which he tests it, the words and images he reaches for that another teacher would not, the failures he recognises before he can explain them, and the part he says cannot be conveyed at all. The last block of each session put the system's own output in front of him and recorded what he said about it. We worked in parallel from three book manuscripts of his, none of them yet published: *Why Write A Script*, *The Power of Cinema Language* and *The Origins of Moving Pictures*.

Encode. The transcripts became prompt modules and a curriculum rather than a summary. The encoding produces three things: a specification of his voice, an evaluation rubric with anchored descriptors written from his own statements of what is good and what is missing, and his worked diagnoses of scenes as the anchors for that rubric.

Calibrate. He reviewed the working system in two recorded sessions, on 7 and 28 March 2026, and the corrections went back into the prompts.

> *CineCoach.paris reflects who I, Paco Wiser, co-author, am through what I have built into it: a space designed to help end-users (students, writers, and others) bring depth, substance, and soul to their stories. I am committed to protecting and fostering what gives their work its true value: originality and unique emotional weight.*
>
> *What I wanted to avoid was telling users how they must write or write in their place. On CineCoach.paris, they will find no dogmatic manuals or ready-made formulas meant to smooth over the rough edges of their stories. My goal is neither to standardize creativity nor to confine users' imagination within rigid rules that might discourage them.*
>
> *Instead, CineCoach.paris offers a rigorous yet supportive framework designed to fuel users' creative fire, help them push past their doubts, and encourage them to bring powerful, unique, and deeply vibrant worlds to the screen, rooted in their own experience, creativity, and imagination.*
>
> – Paco Wiser, co-author, in his own words, written for this paper in August 2026.

5.2 The system

CineCoach is a web application for screenwriting and cinematography students. As the system stood on 28 August 2026 the encoded method had five parts: a four-pass diagnostic (spectator, diagnostician, cinematographer, teacher), a cascade of five questions applied in order ("What is it about? Is there a theme? Does the drama serve the theme? Is it a cliché? Is there emotion?"), an eight-dimension rubric with anchored descriptors, thirteen named reference points from film history, and a five-level curriculum of twenty-one exercises with five gates.

5.3 Where the evidence is applied

We built the system from one teacher's practice between February and March 2026 and we reviewed the evidence afterwards, in August. Each design decision below is therefore consistent with a finding from Sections 3 and 4 rather than derived from it, and Section 5.4 states where the system is not yet consistent.

The student generates first. No surface returns feedback without a submission – a scene of at least 50 characters, an exercise of at least 20, a gate submission of at least 50. This is consistent with Section 3.2 – the student generates, every time.

The system questions and withholds. The teaching-philosophy module states: "Socratic, not didactic. You guide through questions, not lectures. Pull the student toward their own discovery." The curriculum instructs the model to phrase any suggestion as a one-sentence possibility and to

"never write their script or story for them", and to return no suggestions at all when every dimension is strong: "a strong submission needs no hand-holding". A further instruction reads: "Do NOT make feedback so complete that the writer never struggles." This is consistent with Section 4 and Bjork's desirable difficulties.

Feedback is task-level and ends in action. Every evaluation returns per-dimension written feedback and a next step ("End with the path forward. Where does the rewrite begin?"). Suggestions, when given, appear only where a dimension is weak. This is consistent with Section 3.4.

The student is sent into their own experience. The Explore module opens every film conversation by asking for an emotional memory before any analysis, and the first curriculum level asks for "a confession, not a pitch". This is consistent with Madore et al. (2015).

Scaffolding changes with the student. The feedback voice is calibrated by level, from "find the spark" at level one to "peer-to-peer" at level four, and a learner profile with running averages and trends per dimension is carried into every evaluation. This is consistent with Section 3.6, with the limit noted below.

Diversity is designed in. The system runs in fifteen languages. Eleven are official EU languages (Czech, Dutch, English, French, German, Hungarian, Italian, Polish, Portuguese, Romanian and Spanish). The other four are Russian, Turkish, Latin-American Spanish and Brazilian Portuguese. Each of the fifteen carries its own national-cinema curation of 24 films, 360 in total, analysed natively in that language rather than translated, so that a Hungarian student meets Hungarian cinema (Saul fia, say) in Hungarian and a student in Latin America meets Latin-American cinema in their own Spanish. The model is instructed to answer entirely in the student's language. This is a design response to Agarwal et al. (2025) and Atari et al. (2023), not yet a measured one.

5.4 Limits

Our usage data are small – a few dozen real users at the time of writing – and cannot support outcome claims. No published study, ours included, tests a questioning AI on a creative-production task with a delayed, unaided outcome. In September 2026 the administrators who coordinate EICAR’s student film productions received accounts so that they can reach students, and they have asked to test the system themselves. A study with informed consent is being designed from that starting point. At the time of writing the product has no research-consent flow, and one must be added before any such study can run.

The explicit Socratic instruction is wired into the Explore module. The script-evaluation, curriculum and gate modules are diagnostic – they score, explain and end with a question, which is

the high-information feedback of Section 3.4 rather than pure dialogue. The script-evaluation surface currently displays numeric per-dimension scores and a radar chart, which Section 3.3 suggests should be shown after the comments, not before. The suggestion lists, though anchored in the student's text, are the feature closest to Kumar et al.'s strategy lists, and we are converting them to questions that quote the student's own lines. These changes are in progress.

The European AI Act, Regulation (EU) 2024/1689, bears on the design directly. Article 50(1) requires a provider to ensure that a person is told they are interacting with an AI system, and it has applied since 2 August 2026. Article 50(2) requires machine-readable marking of synthetic audio, image, video or text. The scene the student keeps is their own work, and nobody needs to mark it as synthetic. The coach's own output (its questions, its per-dimension comments, its suggestions) is AI-generated text. We do not claim that the exemption for assistive functions removes it from Article 50(2). Whether it does is a legal question, and the text is disclosed as AI-generated in any case under Article 50(1). Point 3(b) of Annex III classifies as high-risk an AI system "intended to be used to evaluate learning outcomes, including when those outcomes are used to steer the learning process" in an educational or vocational training institution. Our eight-dimension rubric evaluates a student's work inside a film school, so if a school were to use its output in assessment or to decide what a student studies next, the system would fall inside that point, and saying that we only ask questions would not settle it. Those obligations were deferred from 2 August 2026 to 2 December 2027 by Regulation (EU) 2026/1744. Article 4, on AI literacy, has applied since 2 February 2025 and binds deployers as well as providers, which places a duty on the school as much as on us. We state this because it is the clearest illustration of our own argument: which of the two designs a system implements decides its legal category, and so a public authority can act on the design.

The system runs on a United States foundation model with a second US model as fallback. The European contribution is the expertise layer, the languages, the curations and the governance. The base model is replaceable and we intend to test European alternatives. We claim sovereignty over the layer that carries the culture (the method, the languages, the curations, the data governance), and we claim it is achievable now.

The encoding is our interpretation. Studies of film and design assessment show that expert judgement is partly tacit (Florén et al. 2025; Dannels and Martin 2008), and our own protocol is not finished either. It calls for a corpus of examples annotated by the expert himself, and that corpus does not yet exist – the anchors in the rubric are our transcriptions of his spoken diagnoses, not cases he has marked up. He has reviewed the system twice. The protocol expects that loop to run for as long as the system does. The tool supplies the specific, actionable part of expert feedback. The master still teaches.

6. Recommendations for the European strategy

The Commission's three pillars for the coming strategy are innovation and competitiveness, fair and ethical models that protect creation and diversity and equipping the sectors for the transition. Each has a concrete instrument in the evidence above.

1. Fund the design, not the technology. "Complement, not replace" is a property of the interface and the difference is measurable (Table 2). European developers can compete on the augmenting design, which general-purpose generators are not built for. We do not propose an eligibility bar. Many tools the sectors need (dubbing, subtitling, visual effects, restoration, previsualisation) produce the artefact, and the evidence for the design distinction is, so far, one pre-registered experiment. We propose two lighter instruments. First, for artist-facing tools in education and training, Creative Europe, Horizon Europe and Digital Europe calls should score whether the system augments the artist's own generation or produces the artefact for them, and should weight that score. Second, every funded creative-AI system should state in its application which of the two designs it implements, so that the choice is visible to the funder and to the user. Our own system would score well on the first instrument, and the reader should weigh this recommendation with that interest in mind. This is consistent with the strategy's own orientation that AI should complement and not replace human creativity and agency. Because the design can be costly to demonstrate, calls should pair the criterion with shared compute or data-access support so that small developers and cultural organisations, the actors the strategy itself flags as most exposed to unequal AI resources, can meet it on equal terms with larger vendors. This recommendation complements, rather than substitutes for, the parallel review of the licensing and rightsholder questions under the CDSM Directive.

2. Measure diversity as an outcome. The homogenisation effect is small in any one study ($d = 0.33$) and compounds across a population that uses the same models. Publicly funded creative AI should be required to report the diversity of outputs across users and across languages (next to adoption and productivity) and the results should be published. Cultural and linguistic diversity then becomes a monitored outcome of AI policy rather than a stated value. This reporting can run through infrastructure the strategy is already building: the EU Cultural Data Hub, the AI Observatory announced in the Apply AI strategy, and the Feedback to Policy exercises conducted by EACEA and REA, rather than requiring a new monitoring mechanism.

3. Put the tools under the schools. Artist-facing AI in film schools, conservatoires and studios should be funded through the institutions – with faculty control of the pedagogy and plain disclosure to students of what the tool will and will not do for them. A licence to a general-purpose

product transfers the design decision (and with it the diversity outcome) to the vendor. Institution-led control also serves the strategy's goal of broader access to diverse cultural content and knowledge for young people, since pedagogy stays answerable to the school rather than to a vendor's product roadmap.

4. Teach the difference. AI literacy for artists (the third pillar's stated aim) must include the distinction in Table 2. An artist who knows that a generating tool raises their output and lowers their ownership, and that a questioning tool does the reverse, can choose, and choosing is the agency the strategy sets out to protect. This distinction, generating versus questioning, is a natural candidate for the transparency obligations the European AI Act already places on providers, giving artist-facing disclosure a legal anchor rather than resting on literacy training alone.

5. Build the evidence. The gap in the literature is precise – no controlled study of a questioning AI on a creative-production task with a delayed unaided outcome in any European language. Such a study would also test the hypothesis left open in Section 4, that questions anchored in the person's own material avoid the loss of unaided skill that generic strategy lists produced. It is a cheap study to fund and the sectors' own schools can run it. We are running one and would rather it be one of many.

6. Anchor sovereignty in governance, not in the model. The strategy names over-reliance on non-EU technology as a risk to European digital cultural sovereignty. European foundation models exist, Mistral AI in Paris among them, and we intend to test them. But a model-residency requirement imposed today would exclude many working European tools, including ours, and it would tie a cultural policy to whichever company trains the weights. Sovereignty should instead be defined and required at the layer that is genuinely European and non-substitutable: the curation of source material, the pedagogy encoded in the system, the language coverage, and the governance of data and consent. Funding criteria should state this explicitly, so that European developers are not forced to choose between compliance and using the best available model.

7. Conclusion

AI is a tool and a tool of unusual power. Used wrongly it will reduce human creativity – the evidence for that is already in (in meta-analysis, in field data from four million artworks and in examination rooms). Used rightly it will augment and accelerate it – and that evidence is in too. Films from Europe have won most of the Academy Awards for international features and most of the Palmes d'Or. What it needs from AI is the making of filmmakers and composers and designers in all 24 of its languages. CineCoach is one European example of that design (built in France from

a French master's teaching). There should be many. Many, because the model inside each one is a commodity and replaceable; what is not replaceable is the curation, the languages and the governance built around it, and that is where Europe's sovereignty in this technology actually lives, not in which company trained the weights. Because the difference between the two designs is a property of the interface it can be specified in a call, required in a contract, taught in a school and measured in a cohort. Diversity of output, not only quality, is one of the things a cohort can be measured on, which means it can be required as a funding condition and published as an outcome, rather than asserted as a hope. A strategy that complements rather than replaces human creativity can begin with the next funding round. This September EICAR in Paris began working with such a machine, starting with the staff who coordinate its student productions. What its students make of it is the evidence that matters next.

Data and materials. The evidence map (87 entries with DOI, verification status and statistics as stated, updated on 5 September 2026 with the corrections of the 28 August verification pass) is deposited alongside this paper. The CineCoach feature inventory with file references is section 8 of that map. **Funding.** None. **Conflict of interest.** The authors founded and own DECCS, which develops CineCoach. **AI use statement.** Literature scans and drafting were assisted by a large language model under the authors' direction. Every statistic was checked against the opened source twice, by the scan and then by a second verification pass under the authors' direction, and is traceable in the deposited map.

References

Academy of Motion Picture Arts and Sciences winners, as compiled in "List of countries by number of Academy Awards for Best International Feature Film", Wikipedia, accessed 28 August 2026. Counts summed from the by-country table, which records country credits rather than awards: the 1949 award shared by Italy and France appears twice, so 79 credits correspond to 78 awards given. The European subset is the 27 present Member States of the European Union together with their predecessor states in the table (West Germany, East Germany and Czechoslovakia), giving 52 wins and 213 nominations. Counting the present 27 alone gives 49 and 198. The compiled table was used because it carries the country credits in one place; the primary record is the Academy's own awards database, which the compilation cites. Other defensible rules give larger figures: any European country with the Soviet Union and Russia excluded gives 57 and 239, and including them gives 61 and 255.

Cannes Film Festival, "Palme d'Or" winners list, as compiled in Wikipedia, accessed 28 August 2026. Share counted from the production-country column of 102 award entries, 1946–2026 (105 table rows less the cancelled festivals of 1939, 1968 and 2020), counting a film once if any credited production country qualifies. On the rule used here — the 27 present Member States together with West Germany and Czechoslovakia — the figure is 54; the present 27 alone give 53, and adding Yugoslavia gives 56. The primary record is the Festival's own archive, which the compilation cites. France alone is credited on 29, which matches the article's own wins-by-country table. Wider rules give larger figures: any European country with Turkey and the Soviet Union excluded gives 66, and including them 69.

Deloitte (2023). *Global Powers of Luxury Goods 2023*. https://www.deloitte.com/global/en/industries/consumer/analysis/gx-cb-global-powers-of-luxury-goods.html

European Commission, DG GROW (n.d.). Cultural and creative industries. https://single-market-economy.ec.europa.eu/sectors/cultural-and-creative-industries_en (accessed 28 August 2026).

European Commission (2025). *Special Eurobarometer 562: Europeans' attitudes towards culture*. Fieldwork February–March 2025. https://doi.org/10.2766/4514147

European Commission (2026). *Call for evidence for an initiative (without an impact assessment): AI strategy for the cultural and creative sectors*. Directorate-General for Education, Youth, Sport and Culture, Unit D1. PLAN/2026/867, Ref. Ares(2026)7389748, 27 July 2026. Feedback open to 11 September 2026. https://ec.europa.eu/info/law/better-regulation/have-your-say/initiatives/18592-Artificial-intelligence-strategy-for-the-cultural-and-creative-sectors_en

European Union (n.d.). Countries. https://european-union.europa.eu/principles-countries-history/eu-countries_en (accessed 5 September 2026).

European Union (n.d.). Languages. https://european-union.europa.eu/principles-countries-history/languages_en (accessed 28 August 2026).

Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act). *OJ L, 2024/1689, 12.7.2024.* https://eur-lex.europa.eu/eli/reg/2024/1689/oj

Regulation (EU) 2026/1744 of the European Parliament and of the Council of 8 July 2026 amending Regulations (EU) 2024/1689, (EU) 2018/1139 and (EU) 2023/1230 as regards the simplification of the implementation of harmonised rules on artificial intelligence (Digital Omnibus on AI). *OJ L, 2026/1744, 24.7.2026.* https://eur-lex.europa.eu/eli/reg/2026/1744/oj

Eurostat (2026). Culture statistics: cultural employment. Data extracted June 2026. https://ec.europa.eu/eurostat/statistics-explained/index.php?title=Culture_statistics_-_cultural_employment

Abrami, P. C., Bernard, R. M., Borokhovski, E., Waddington, D. I., Wade, C. A., & Persson, T. (2015). Strategies for teaching students to think critically: A meta-analysis. *Review of Educational Research, 85*(2). https://doi.org/10.3102/0034654314551063

Adesope, O. O., Trevisan, D. A., & Sundararajan, N. (2017). Rethinking the use of tests: A meta-analysis of practice testing. *Review of Educational Research, 87*(3), 659–701. https://doi.org/10.3102/0034654316689306

Agarwal, D., Naaman, M., & Vashistha, A. (2025). AI suggestions homogenize writing toward Western styles and diminish cultural nuances. *CHI 2025*. https://doi.org/10.1145/3706598.3713564

Alfieri, L., Brooks, P. J., Aldrich, N. J., & Tenenbaum, H. R. (2011). Does discovery-based instruction enhance learning? *Journal of Educational Psychology, 103*(1). https://doi.org/10.1037/a0021017

Amabile, T. M. (1979). Effects of external evaluation on artistic creativity. *Journal of Personality and Social Psychology, 37*, 221–233. https://doi.org/10.1037/0022-3514.37.2.221 (primary not opened; reported via Amabile, 2012.)

Amabile, T. M. (1985). Motivation and creativity: Effects of motivational orientation on creative writers. *Journal of Personality and Social Psychology, 48*(2), 393–399. https://doi.org/10.1037/0022-3514.48.2.393

Amabile, T. M. (2012). Componential theory of creativity. *Journal of Creative Behavior*. https://doi.org/10.1002/jocb.001

Amabile, T. M., Hennessey, B. A., & Grossman, B. S. (1986). Social influences on creativity: The effects of contracted-for reward. *Journal of Personality and Social Psychology, 50*(1), 14–23.

Anderson, B. R., Shah, J. H., & Kreminski, M. (2024). Homogenization effects of large language models on human creative ideation. *Creativity & Cognition 2024*. arXiv:2402.01536

Atari, M., Xue, M. J., Park, P. S., Blasi, D., & Henrich, J. (2023). Which humans? *PsyArXiv* (preprint). https://osf.io/preprints/psyarxiv/5b26t

Barbieri, C. A., Miller-Cotto, D., Clerjuste, S. N., & Chawla, K. (2023). A meta-analysis of the worked examples effect on mathematics performance. *Educational Psychology Review, 35*(1). https://doi.org/10.1007/s10648-023-09745-1

Bastani, H., Bastani, O., Sungu, A., Ge, H., Kabakcı, Ö., & Mariman, R. (2025). Generative AI without guardrails can harm learning: Evidence from high school mathematics. *PNAS*. https://doi.org/10.1073/pnas.2422633122

Beaty, R. E., Benedek, M., Silvia, P. J., & Schacter, D. L. (2016). Creative cognition and brain network dynamics. *Trends in Cognitive Sciences, 20*(2), 87–95. https://doi.org/10.1016/j.tics.2015.10.004

Beaty, R. E., Benedek, M., Wilkins, R. W., et al. (2014). Creativity and the default network. *Neuropsychologia, 64*, 92–98. https://doi.org/10.1016/j.neuropsychologia.2014.09.019

Beaty, R. E., Kenett, Y. N., Christensen, A. P., et al. (2018). Robust prediction of individual creative ability from brain functional connectivity. *PNAS, 115*(5), 1087–1092. https://doi.org/10.1073/pnas.1713532115

Benedek, M., Jauk, E., Fink, A., et al. (2014). To create or to recall? Neural mechanisms underlying the generation of creative new ideas. *NeuroImage, 88*, 125–133. https://doi.org/10.1016/j.neuroimage.2013.11.021

Bertsch, S., Pesta, B. J., Wiscott, R., & McDaniel, M. A. (2007). The generation effect: A meta-analytic review. *Memory & Cognition, 35*(2), 201–210. https://doi.org/10.3758/BF03193441

Bisra, K., Liu, Q., Nesbit, J. C., Salimi, F., & Winne, P. H. (2018). Inducing self-explanation: A meta-analysis. *Educational Psychology Review, 30*, 703–725. https://doi.org/10.1007/s10648-018-9434-x

Bjork, R. A. (1994). Memory and metamemory considerations in the training of human beings. In J. Metcalfe & A. Shimamura (Eds.), *Metacognition: Knowing about knowing* (pp. 185–205). MIT Press.

Blasco, A., & Charisi, V. (2025). AI chatbots in K-12 education: An experimental study of Socratic vs. non-Socratic AI agents and the role of step-by-step reasoning. SSRN working paper 5040921, version of 22 November 2025 (first posted 4 December 2024, last revised 26 November 2025). https://doi.org/10.2139/ssrn.5040921 (preprint, not peer reviewed; the SSRN record still carries the earlier title, "...non-Socratic approaches...").

Bloom, B. S. (1984). The 2 sigma problem. *Educational Researcher, 13*(6), 4–16. https://doi.org/10.3102/0013189X013006004

Butler, R. (1988). Enhancing and undermining intrinsic motivation: The effects of task-involving and ego-involving evaluation on interest and performance. *British Journal of Educational Psychology, 58*, 1–14. https://doi.org/10.1111/j.2044-8279.1988.tb00874.x

Byron, K., & Khazanchi, S. (2012). Rewards and creative performance: A meta-analytic test. *Psychological Bulletin, 138*(4), 809–830. https://doi.org/10.1037/a0027652

Byron, K., Khazanchi, S., & Nazarian, D. (2010). The relationship between stressors and creativity: A meta-analysis. *Journal of Applied Psychology, 95*(1), 201–212. https://doi.org/10.1037/a0017868

Chen, O., Kalyuga, S., & Sweller, J. (2017). The expertise reversal effect is a variant of the more general element interactivity effect. *Educational Psychology Review, 29*. https://doi.org/10.1007/s10648-016-9359-1

Cogdell-Brooke, L. S., Sowden, P. T., Violante, I. R., & Thompson, H. E. (2020). A meta-analysis of fMRI studies of divergent thinking using activation likelihood estimation. *Human Brain Mapping*. https://doi.org/10.1002/hbm.25170

Dannels, D. P., & Martin, K. N. (2008). Critiquing critiques: A genre analysis of feedback across novice to expert design studios. *Journal of Business and Technical Communication, 22*(2), 135–159. https://doi.org/10.1177/1050651907311923

de Jesus, S. N., Rus, C. L., Lens, W., & Imaginário, S. (2013). Intrinsic motivation and creativity related to product: A meta-analysis. *Creativity Research Journal, 25*(1), 80–84. https://doi.org/10.1080/10400419.2013.752235

de Rooij, A., & Biskjaer, M. M. (2026). Does generative AI make us think alike? A systematic review and meta-analysis of homogenization effects in human–AI co-creation. *ECCE 2026* (preprint). https://doi.org/10.31234/osf.io/rz5s4_v1

Deci, E. L., Koestner, R., & Ryan, R. M. (1999). A meta-analytic review of experiments examining the effects of extrinsic rewards on intrinsic motivation. *Psychological Bulletin, 125*(6), 627–668. https://doi.org/10.1037/0033-2909.125.6.627

Dietrich, A., & Kanso, R. (2010). A review of EEG, ERP, and neuroimaging studies of creativity and insight. *Psychological Bulletin, 136*(5), 822–848. https://doi.org/10.1037/a0019749

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

Dunlosky, J., Rawson, K. A., Marsh, E. J., Nathan, M. J., & Willingham, D. T. (2013). Improving students' learning with effective learning techniques. *Psychological Science in the Public Interest, 14*(1), 4–58. https://doi.org/10.1177/1529100612453266

Ericsson, K. A., & Harwell, K. W. (2019). Deliberate practice and proposed limits on the effects of practice on the acquisition of expert performance. *Frontiers in Psychology*. https://doi.org/10.3389/fpsyg.2019.02396

Fan, Y., Tang, L., Le, H., et al. (2025). Beware of metacognitive laziness: Effects of generative artificial intelligence on learning motivation, processes, and performance. *British Journal of Educational Technology, 56*(2), 489–530. https://doi.org/10.1111/bjet.13544

Florén, M., Bezemer, J., & Domingo, M. (2025). Assessing film in higher education: Straddling academic and professional conventions. *Learning, Media and Technology*. https://doi.org/10.1080/17439884.2025.2471405

Gerlich, M. (2025). AI tools in society: Impacts on cognitive offloading and the future of critical thinking. *Societies, 15*(1), 6. https://doi.org/10.3390/soc15010006

Hattie, J., & Timperley, H. (2007). The power of feedback. *Review of Educational Research, 77*(1), 81–112. https://doi.org/10.3102/003465430298487

Holzner, N., Maier, S., & Feuerriegel, S. (2025). Generative AI and creativity: A systematic literature review and meta-analysis. arXiv:2505.17241 (preprint).

Hopkins, E. J., Weisberg, D. S., & Taylor, J. C. V. (2016). The seductive allure is a reductive allure. *Cognition, 155*, 67–76. https://doi.org/10.1016/j.cognition.2016.06.011

Hou, Y., Wang, Y., Wang, C., Wang, X., & Yang, X. (2025). The double-edged roles of generative AI in the creative process: Experiments on design work. *Information Systems Research*. https://doi.org/10.1287/isre.2024.0937

Kalyuga, S., Ayres, P., Chandler, P., & Sweller, J. (2003). The expertise reversal effect. *Educational Psychologist, 38*(1). https://doi.org/10.1207/S15326985EP3801_4

Kao, S., Grant, P., & Woltering, S. (2025). Socratic AI in K–12 science classrooms. Research Square (preprint). https://doi.org/10.21203/rs.3.rs-8118546/v1

Kestin, G., Miller, K., Klales, A., Milbourne, T., & Ponti, G. (2025). AI tutoring outperforms in-class active learning. *Scientific Reports, 15*. https://doi.org/10.1038/s41598-025-97652-6

Kirschner, P. A., Sweller, J., & Clark, R. E. (2006). Why minimal guidance during instruction does not work. *Educational Psychologist, 41*(2). https://doi.org/10.1207/s15326985ep4102_1

Kluger, A. N., & DeNisi, A. (1996). The effects of feedback interventions on performance. *Psychological Bulletin, 119*(2), 254–284. https://doi.org/10.1037/0033-2909.119.2.254

Koestner, R., Ryan, R. M., Bernieri, F., & Holt, K. (1984). Setting limits on children's behavior: The differential effects of controlling vs. informational styles on intrinsic motivation and creativity. *Journal of Personality, 52*(3). https://doi.org/10.1111/j.1467-6494.1984.tb00879.x

Kosmyna, N., et al. (2025). Your brain on ChatGPT: Accumulation of cognitive debt when using an AI assistant for essay writing task. arXiv:2506.08872 (preprint).

Kulik, J. A., & Fletcher, J. D. (2016). Effectiveness of intelligent tutoring systems: A meta-analytic review. *Review of Educational Research, 86*(1). https://doi.org/10.3102/0034654315581420

Kumar, H., Vincentius, J., Jordan, E., & Anderson, A. (2024). Human creativity in the age of LLMs: Randomized experiments on divergent and convergent thinking. arXiv:2410.03703 (preprint).

Lazonder, A. W., & Harmsen, R. (2016). Meta-analysis of inquiry-based learning: Effects of guidance. *Review of Educational Research, 86*(3), 681–718. https://doi.org/10.3102/0034654315627366

Liu, Z., Zuo, H., & Lu, Y. (2025). The impact of ChatGPT on students' academic achievement: A meta-analysis. *Journal of Computer Assisted Learning*. https://doi.org/10.1111/jcal.70096 (abstract opened).

Macnamara, B. N., Hambrick, D. Z., & Oswald, F. L. (2014). Deliberate practice and performance in music, games, sports, education, and professions: A meta-analysis. *Psychological Science, 25*(8), 1608–1618. https://doi.org/10.1177/0956797614535810 Corrigendum (2018): https://doi.org/10.1177/0956797618769891

Madore, K. P., Addis, D. R., & Schacter, D. L. (2015). Creativity and memory: Effects of an episodic-specificity induction on divergent thinking. *Psychological Science, 26*(9), 1461–1468. https://doi.org/10.1177/0956797615591863

Maier, S., Schneider, J., & Feuerriegel, S. (2025). Partnering with generative AI: Experimental evaluation of human-led and model-led interaction in human-AI co-creation. arXiv:2510.23324 (preprint).

McCurdy, M. P., Viechtbauer, W., Sklenar, A. M., Frankenstein, A. N., & Leshikar, E. D. (2020). Theories of the generation effect and the impact of generation constraint: A meta-analytic review. *Psychonomic Bulletin & Review*. https://doi.org/10.3758/s13423-020-01762-3

Rowland, C. A. (2014). The effect of testing versus restudy on retention: A meta-analytic review of the testing effect. *Psychological Bulletin, 140*(6), 1432–1463. https://doi.org/10.1037/a0037559

Scott, G., Leritz, L. E., & Mumford, M. D. (2004). The effectiveness of creativity training: A quantitative review. *Creativity Research Journal, 16*(4), 361–388. https://doi.org/10.1080/10400410409534549

Seli, P., Ragnhildstveit, A., Orwig, W., Bellaiche, L., Spooner, S., & Barr, N. (2025). Beyond the brush: Human versus artificial intelligence creativity in the realm of generative art. *Psychology of Aesthetics, Creativity, and the Arts*. https://doi.org/10.1037/aca0000743

Sinha, T., & Kapur, M. (2021). When problem solving followed by instruction works: Evidence for productive failure. *Review of Educational Research, 91*(5). https://doi.org/10.3102/00346543211019105

Sio, U. N., & Ormerod, T. C. (2009). Does incubation enhance problem solving? A meta-analytic review. *Psychological Bulletin, 135*(1), 94–120. https://doi.org/10.1037/a0014212

Slamecka, N. J., & Graf, P. (1978). The generation effect: Delineation of a phenomenon. *Journal of Experimental Psychology: Human Learning and Memory, 4*(6), 592–604.

Sowden, P. T., Pringle, A., & Gabora, L. (2015). 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

Thompson, D. (2018). Lights, camera, action research! Engaging filmmaking students in feedback. *Compass: Journal of Learning and Teaching*. https://journals.gre.ac.uk/index.php/compass/article/view/711

VanLehn, K. (2011). The relative effectiveness of human tutoring, intelligent tutoring systems, and other tutoring systems. *Educational Psychologist, 46*(4), 197–221. https://doi.org/10.1080/00461520.2011.611369

Wang, J., & Fan, W. (2025). The effect of ChatGPT on students' learning performance, learning perception, and higher-order thinking: Insights from a meta-analysis. *Humanities and Social Sciences Communications, 12*, 621. https://doi.org/10.1057/s41599-025-04787-y. Retracted 22 April 2026 for discrepancies in the meta-analysis: https://doi.org/10.1057/s41599-026-07310-z

Wang, Y., Kim, S., Peng, Y., & Wang, Z. (2025). Exploring creativity in human–AI co-creation: A comparative study across design experience. *Frontiers in Computer Science*. https://doi.org/10.3389/fcomp.2025.1672735

Weisberg, D. S., Keil, F. C., Goodstein, J., Rawson, E., & Gray, J. R. (2008). The seductive allure of neuroscience explanations. *Journal of Cognitive Neuroscience, 20*(3), 470–477. https://doi.org/10.1162/jocn.2008.20040

Wisniewski, B., Zierer, K., & Hattie, J. (2020). 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

Wu, X., Zhu, P., Zhang, J., Yin, M., & Wang, Y. (2026). ChatGPT's impact on student learning outcomes: A meta-analysis of 35 experimental studies. *Humanities and Social Sciences Communications, 13*, 684. https://doi.org/10.1057/s41599-026-07019-z

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