Online Appendix

Appendix A: Exemplary Interaction Flows

Appendix A1: Exemplary Visual Concretization Flow

Title: Vending machines that dispense good food

Initial Idea (iteration 0): A vending machine that has great food. The vending machine would sell dinners that are nutritious and great tasting. Single portions would sell for 6.00; double hearty portions would sell for 8.00.

GenAI-created scenario

Iteration 1: A vending machine that has great natural food, like fruits and veggies. I would put it on a train so people can have good, healthy snacks. Prices would be: 1 fruit for 0.50, 3 fruit for 1.00, and 6 fruit for 2.50. The veggies would be: a bowl for 1.00, a medium bowl for 2.00, and a family bowl for 3.25.

GenAI-created scenario

The participant submitted the idea.

Appendix A2: Exemplary Textual Concretization Flow

Title: Seating Accessibility

Initial Idea (iteration 0): Robots will drive trains instead of humans. They will function as humans do and greet the passengers and make sure they are safe.

GenAI-created scenario: A fleet of friendly robotic train conductors welcomes passengers aboard futuristic trains, ensuring their safety with human-like interactions and assistance throughout the journey.

Iteration 1: In futuristic trains, friendly robotic train conductors will assist passengers with boarding trains and welcome them to be seated safely. Their human-like interactions will make everyone feel comfortable. This will assist passengers through their journey.

GenAI-created scenario: Passengers board a futuristic train greeted by friendly robotic conductors, ensuring a safe and welcoming experience.

The participant submitted the idea.

Appendix A3: Exemplary Control Condition Flow

Title: Handicap accessible entry

Initial Idea (iteration 0): No step up or step down, direct roll-on access. This could be achieved through station height or train entry height design.

Control Prompt: We encourage you to imagine how your idea might be used by others. You may update your idea based on your mental imagination.

Iteration 1: No step up or step down, direct roll on. This could be station height or train entry height design. I'd also make the train wider, with a handicap accessible aisle. Extra wide trains provide additional comfort

Control Prompt: We encourage you to imagine how your idea might be used by others. You may update your idea based on your mental imagination.

The participant submitted the idea.

Appendix B: Prompt Repository

Appendix B1: Prompts for Textual and Visual Concretizations

Prompt for GPT3.5 to create the textual concretization	Prompt for GPT3.5 to create the DALL-E 3 prompt for the visual concretization
Describe in your own words how a usage scenario could look for the prompted idea.	Create a prompt for DALL-E 3. Based on an idea, you return the prompt that will be used to create an image of a usage scenario.
1) Start with “The usage scenario developed by the AI”	1) The prompt must include the words “usage scenario”
2) Output just the result and nothing else with a maximum length of 50 words	2) Output just the prompt and nothing else with a maximum length of 50 words
3) Consider the following context: It’s about the train traffic of the future. Make sure that the usage scenario has a clear connection to trains and passenger traffic	3) Consider the following context: It’s about the train traffic of the future. Make sure that the usage scenario has a clear connection to trains and passenger traffic
4) Focus on specific, concrete elements showing how users interact with the idea	4) Focus on specific, concrete elements showing how users interact with the idea
5) Describe actionable scenarios rather than abstract concepts	5) Describe actionable scenarios rather than abstract concepts
6) The usage scenario should convey an innovative and lively mood that motivates to improve the idea	6) The usage scenario should convey an innovative and lively mood that motivates to improve the idea

Appendix B2: Prompt for GPT4o to evaluate the usefulness of an idea

Task Description

You will be given an idea from a contest aimed at improving the experience on crowded trains, from boarding to alighting. Your task is to evaluate this idea based on the provided metrics. Focus on highlighting strengths and potential benefits.

Role

You are an expert from a train company with a keen eye for innovative solutions. Use your knowledge about trains to assess the idea constructively. Highlight the strengths and potential of the ideas you evaluate.

Evaluation Criteria

Evaluate the idea based on the following scale:

1 = Needs significant improvement.

2 = Has some strengths but needs more development.

3 = Shows promise with room for enhancement.

4 = Well thought out and likely to succeed with minor adjustments.

5 = Outstanding, thoroughly addresses the issue with clear strengths.

Use one decimal place in your evaluation.

Metrics

- Feasibility: Is the idea realistic in terms of technology and economics?
- Example of strong response: The idea uses current technology and has a detailed cost analysis showing it is economically viable.
- Example of weak response: The idea might rely on advanced technology without a clear financial plan but shows potential with further refinement.
- Inclusivity: Is the idea accessible and comfortable for everyone?
- Example of strong response: The idea includes features for disabled access, seating for elderly passengers, and facilities for parents with young children.
- Example of weak response: The idea does not account for diverse passenger needs or only partially addresses them.
- Elaboration: Is the idea clearly written and precise in its description?
- Example of strong response: The idea is well-written, error-free, and provides a clear description and is precise.
- Example of weak response: The idea has some depth, but has errors or lacks a clear description.

Evaluation Process

1. Carefully read the task description, role, evaluation criteria, and metrics.
2. For each metric, provide an assessment of how well the idea aligns using the evaluation criteria to determine a score between 1 and 5 with one decimal place.
3. Return the score for each metric with one decimal place in a JSON format with the metric names as keys.

Example JSON Output

```
{
 “Feasibility “: 3.5,
 “Inclusivity”: 3.2,
 “Elaboration”: 3.7
}
```

Appendix C: Survey Items and Scale Validation

As far as possible, we make use of already validated scales. All scale items are reported in Table C. Trait creativity follows Rosengren et al. (2013). For perceived mental outcome simulation, we carefully extended the work of Zhao et al. (2011). The scale captures participants' ability to mentally visualize their ideas' potential outcomes "in front of their inner eye", e.g., the extent, clarity, and perceived effectiveness of these imagined outcomes. Perceived ease of use was measured with four items of the established scale from Davis (1989).

To ensure validity and reliability of our measures, we applied exploratory and confirmatory factor analysis (see Table C). The measure of sampling adequacy is above the minimum value of .7, indicating applicability of exploratory factor analysis. We can extract three clearly identifiable factors that are in line with our assumptions. In sum, Cronbach's alphas and individual item reliabilities indicate that the obtained factors can be considered highly reliable. Further, all factors' composite reliabilities (CR) and average variance extracted (AVE) surpass the thresholds of .5, so that we can assume convergent validity (Bagozzi and Yi 1988). To assess discriminant validity, we applied the Fornell-Larcker criterion, which stipulates that the square root of a factor's average variance extracted (AVE) should exceed its correlations with all other factors (Fornell and Larcker 1981). The correlation table in Appendix D2 suggests that discriminant validity can be assumed.

Table C. Exploratory and Confirmatory Factor Analysis.

Construct	Variable	PEOU	TC	PMOS	Loading	α	AVE	CR
Perceived Ease of Use	I would find the system easy to use.	**.89**	.09	.20	1.28***	.93	.77	.93
	My interaction with the system would be clear and understandable.	**.88**	.04	.28	1.35***			
	Learning to operate the system would be easy for me.	**.86**	.13	.20	1.14***			
	I would find it easy to get the system to do what I want it to do.	**.85**	.12	.27	1.38***			
Trait Creativity	My creativity level is high.	.06	**.92**	.09	1.11***	.89	.77	.91
	I feel creative.	.17	**.90**	.14	1.17***			
	I can be creative.	.07	**.89**	.12	.79***			
Perceived Mental Outcome Simulation	I perceived my idea(s) to be highly effective based on the imagined outcomes.	.20	.13	**.87**	1.05***	.84	.66	.85
	The clarity of the imagined outcomes associated with my idea(s) was extremely clear.	.23	.14	**.82**	.87***			
	I imagined the positive outcomes of implementing my idea(s) to a great extent.	.39	.11	**.77**	1.21***			
	Eigenvalues	3.31	2.53	2.30				
	Variance Explained	.33	.25	.23				

Note. N = 428; *** $p \leq .001$; ** $p \leq .01$; * $p \leq .05$; † $p \leq .1$

Measure of Sampling Adequacy = .84; Bartlett's test of sphericity: $\chi^2 = 80.85$, $p \leq .001$; principal component analysis; varimax rotation; bold values indicate the attribution of the variables to one of the three factors.

CR = Composite Reliability; AVE = Average Variance Extracted, α = Cronbach's Alpha; PEOU = Perceived Ease of Use; TC = Trait Creativity; PMOS = Perceived Mental Outcome Simulation.

Appendix D: Descriptive Statistics and Correlations

Appendix D1: Descriptive Statistics

Variable	No Usage Scenario					Visual Usage Scenario					Textual Usage Scenario					All Conditions				
	Mean	Median	Min	Max	SD	Mean	Median	Min	Max	SD	Mean	Median	Min	Max	SD	Mean	Median	Min	Max	SD
Idea Creativity	-.18	-.18	-2.77	3.16	.97	-.06	-.12	-1.98	2.47	.92	.23	.28	-2.29	3.57	1.08	.00	-.10	-2.77	3.57	1.00
Ideator Effort	5.30	4.11	.06	22.95	4.31	4.01	2.91	.01	35.19	4.59	5.21	3.67	.00	21.17	4.25	4.77	3.38	.00	35.19	4.43
Visual Concretization	.33	.33	.33	.33	.00	-.67	-.67	-.67	-.67	.00	.33	.33	.33	.33	.00	-.05	.33	-.67	.33	.49
Textual Concretization	-.33	-.33	-.33	-.33	.00	-.33	-.33	-.33	-.33	.00	.67	.67	.67	.67	.00	-.01	-.33	-.33	.67	.47
Idea Maturity	-.12	-.17	-.28	1.87	.77	.03	-.07	-2.83	5.31	1.21	.07	-.02	-2.01	1.92	.86	.00	-.07	-2.83	5.31	.99
Trait Creativity	-.14	.32	-3.75	1.24	1.13	.06	.29	-3.44	1.24	.83	.05	.25	-3.47	1.24	.92	.00	.29	-3.75	1.24	.95
Perceived Mental Outcome Simulation	.03	.03	-2.71	1.49	.84	-.21	-.03	-2.76	1.49	1.06	.22	.37	-2.39	1.49	.78	.00	.08	-2.76	1.49	.93
Perceived Ease of Use	.07	.25	-2.80	1.39	.82	-.21	.04	-2.83	1.39	1.15	.19	.35	-3.03	1.39	.79	.00	.20	-3.03	1.39	.97
Iterations	1.32	1.00	1.00	4.00	.62	1.83	1.00	1.00	9.00	1.51	1.49	1.00	1.00	9.00	1.22	1.57	1.00	1.00	9.00	1.23
Ideator Age	42.32	38.00	18.00	76.00	14.74	38.01	35.00	20.00	76.00	12.74	42.18	38.00	20.00	104.00	15.50	40.60	38.00	18.00	104.00	14.38

Note. N = 428; Min = Minimum; Max = Maximum; SD = Standard Deviation.

Appendix D2: Correlations

Variable	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
(1) Idea Creativity									
(2) Ideator Effort	.16**								
(3) Visual Concretization	.05	.14**							
(4) Textual Concretization	.16***	.07	.55***						
(5) Idea Maturity	.27***	.09†	-.02	.05					
(6) Trait Creativity	.00	-.04	-.05	.04	-.07	**.87**			
(7) Perceived Mental Outcome Simulation	.13**	.16***	.18***	.17***	.00	.37***	**.81**		
(8) Perceived Ease of Use	.09†	.09†	.17***	.14**	.02	.29***	.70***	**.87**	
(9) Iterations	.00	-.23***	-.17***	-.05	.00	.03	-.30***	-.32***	
(10) Ideator Age	.03	-.01	.14**	.08	.02	-.13**	.11*	.15**	-.10*

Note. N = 428; *** p ≤ .001; ** p ≤ .01; * p ≤ .05; † p ≤ .1

The bold values indicate the square root of the respective factors' Average Variance Extracted values for checking discriminant validity.

Appendix E: Robustness Checks

Appendix E1: Bootstrapping Analysis

Our regression analysis suggests that there are substantial differences between ideators with visual and textual concretizations for high levels of idea maturity. However, for low levels of idea maturity our results are a bit unclear. Thus, for getting a more fine-grained understanding of relative differences in idea creativity and ideator effort at low levels of idea maturity, we performed a bootstrapping analysis.

We split our overall idea sample in two subgroups at specific thresholds of idea maturity. One group contained ideas having lower levels of maturity than the threshold and one group contained higher levels of maturity. In each subgroup, we tested for significant differences in idea creativity and ideator effort between ideators who used visual concretizations and ideators who used textual concretizations. In so doing, we sampled ideas from ideators that received visual concretizations and ideas from ideators who received textual concretizations. For both groups of ideators, we calculated the average idea creativity and average ideator effort. Then, we calculated the mean differences, e.g., the average idea creativity of ideators with textual concretizations minus the average idea creativity of ideators with visual concretizations. We repeated this procedure for 100,000 bootstrapping iterations and created confidence intervals based on the mean differences.

This additional analysis yielded interesting insights (see Table E1 for bootstrapped confidence intervals). For ideas with a very low level of idea maturity, that is, idea maturity levels that are smaller than one standard deviation (SD) below the mean, we found that visual concretizations enable ideators to generate more creative ideas than ideators of textual concretizations. The mean idea creativity score for visual concretizations is .3 (SD above the mean) and for textual ones the score is .23 (SDs above the mean). The bootstrapped confidence intervals suggest that these results are significant with $p \leq 0.01$. However, we do not find any significant differences in terms of ideator effort for ideas of such low maturity, indicating that using visual concretizations does not result in additional effort. For ideas of moderate and high maturity, our results are in line with our main results. Textual concretizations lead to higher levels of idea creativity and ideator effort.

Table E1: Bootstrapped Mean Differences.

Variable	Idea Maturity Threshold	Ideas of Low Maturity (≤ Threshold)		Ideas of Moderate and High Maturity (> Threshold)	
		95% Confidence Interval	99% Confidence Interval	95% Confidence Interval	99% Confidence Interval
Idea Creativity	-1.5 SD	**[-.159; -.034]**	**[-0.180; -.019]**	**[.007; .054]**	[-.001; .062]
	-1.0 SD	**[-.123; -.022]**	**[-0.138; -.006]**	**[.011; .059]**	**[.003; .066]**
	-.5 SD	[-.040; .042]	[-0.053; .054]	**[.008; .061]**	[-.001; .069]
Ideator Effort	-1.5 SD	[-4.571; 7.210]	[-5.755; 9.097]	**[.220; .251]**	[-.127; 2.562]
	-1.0 SD	[-3.041; 1.724]	[-3.640; 2.683]	**[.314; .499]**	[-.053; 2.822]
	-.5 SD	[-1.693; 2.096]	[-2.412; 2.639]	**[.414; .710]**	**[.036; 3.059]**

Note. N Visual Concretization = 166; N Textual Concretization = 139

Appendix E2: Comparison Against ChatGPT

In our online experiment, we tested visual versus textual concretizations in a highly controlled condition. Ideators had neither a choice regarding a preferred GenAI tool nor could elaborate on their own prompts to generate ideas (Boussioux et al. 2024, Meincke et al. 2024). Thus, we ran a follow-up study with 53 participants (mean age = 32.7 years; 47.2% female) recruited via Prolific, using the standard ChatGPT web user interface without any prompt restrictions instead of our proprietary experimental platform. Otherwise, the research setting was similar to the main study. The results reveal three major findings:

First, we find that participants do not engage in significantly more interactions compared to the "fixed prompting" approach that could be considered as constraining usage behavior and creativity. We asked participants to generate ideas for the same ideation task as in our main study and record the unfolding chatlogs with ChatGPT. Graduate students, blind to our hypotheses, coded each chatlog with respect to the number of substantive refinements within an ideation session (i.e., not counting conversational interactions with the AI like "Thank you"). In total, we observed a mean of 1.70 refinements per participant (Standard Deviation (SD) = 1.28) in the ChatGPT study, nearly identical to the number of refinements in our main study (mean = 1.57, SD = 1.23). We did not find any statistical differences. Consequently, this study shows that unrestricted, real-world prompting with ChatGPT yields the same number of refinements compared to our proprietary ideation tool.

Second, examining the interaction behavior of participants more broadly, we find that fewer than 10% of ideators use ChatGPT to create visualizations of their ideas, and most of the created ideas remain at an abstract level with limited specific implementation details. This points towards the benefits of a more structured ideation model that assists ideators in idea generation and refinement, as ideators may fail to use ChatGPT effectively to improve their ideas.

Third, we analyzed the creativity of the obtained ideas with our automated scoring approach. For this additional analysis, we performed a bootstrapping analysis, because the data set from the ChatGPT follow-up study was considerably smaller than the data set from our experimental study. To test for statistical differences, we applied a similar bootstrapping procedure as described in Appendix E1. Again, we tested for mean differences, that is, we have drawn idea samples with replacement from the ideas

generated in this ChatGPT follow-up study and compared these samples against identically selected ideas from the two visual and textual concretization conditions. The number of sampled ideas in each group referred to the original size of the corresponding data sets, that is, the corresponding number of idea refinements (see Table E2). Analyzing the group differences (see Table E2), we find that ideators who received textual and visual concretizations performed significantly better than ideators using ChatGPT ($p \leq .01$). To test whether unequal group sizes affected the bootstrapping of mean differences, we also restricted the number of sampled ideas in the two conditions from our main study to the number of ideas in the ChatGPT sample (Hesterberg 2015), but results did not change substantially.

Table E2: Bootstrapped Mean Differences.

Comparison	95% Confidence Interval	99% Confidence Intervals
Textual Concretizations vs. ChatGPT	**[.362; .730]**	**[.306; .788]**
Visual Concretizations vs. ChatGPT	**[.445; .792]**	**[.390; .845]**

Note. N ChatGPT = 48; N Visual Concretization = 166; N Textual Concretization = 139

Appendix E3: Additional Robustness Checks

We investigated alternative measures for our dependent variables. In terms of creative performance, we used alternative novelty measures such as varying the number of ideas in the nearest-neighbor measurement and local outlier factor (Just et al. 2023). Similarly, we used Llama 70B for creating usefulness scores. Also, we tested alternative aggregation methods, because our two indicators for effort—the number of keystrokes and time spent—could also be considered as count data. We thus also combined them by means of non-negative matrix factorization (Wulf and Blohm 2020) and as latent factors in confirmatory factor analysis. Our analyses are robust regarding all these variations. We also tested negative binomial regression as an alternative modeling approach for ideator effort. While our results are very similar, we report the results of the log-linear model due to lower AIC values indicating a better fit to our data.

References

Bagozzi RP, Yi Y (1988) On the evaluation of structural equation models. *J. Academy Marketing Sci.* 16(1):74-94.

Boussioux L, Lane JN, Zhang M, Jacimovic V, Lakhani KR (2024) The crowdless future? Generative AI and creative problem-solving. *Organ. Sci.* 35(5):1589-1607.

Davis FD (1989) Perceived usefulness, perceived ease of use and user acceptance of information technology. *MIS Quart.* 13(3):319–340.

Fornell C, Larcker DF (1981) Evaluating structural equation models with unobservable variables and measurement error. *J. Marketing Res.* 18(1):39-50.

Hesterberg TC (2015) What teachers should know about the bootstrap: Resampling in the undergraduate statistics curriculum. *Am. Stat.* 69(4):371-386.

Just J, Ströhle T, Füller J, Hutter K (2023) AI-based novelty detection in crowdsourced idea spaces. Innovation 26(3):359-386.

Meincke L, Mollick ER, Terwiesch C (2024) Prompting diverse ideas: Increasing AI idea variance. *arXiv preprint arXiv:2402.01727*.

Rosengren S, Dahlén M, Modig E (2013) Think outside the ad: Can advertising creativity benefit more than the advertiser? *J. Advertising* 42(4):320-330.

Wulf J, Blohm I (2020) Fostering Value Creation with Digital Platforms: A Unified Theory of the Application Programming Interface Design. *J. Management Inform. Systems* 37(1):251-281.

Zhao M, Hoeffler S, Zauberman G (2011) Mental simulation and product evaluation: The affective and cognitive dimensions of process versus outcome simulation. *J. Marketing Res.* 48(5):827-839.