# Formulating or Fixating: Effects of Examples on Problem Solving Vary as a Function of Example Presentation Interface Design


Joel Chan
joelchan@umd.edu
College of Information Studies, University of Maryland
USA

Zijian Ding
ding@umd.edu
College of Information Studies, University of Maryland
USA

Eesh Kamrah
kamraheesh@umd.edu
Department of Mechanical Engineering, University of Maryland
USA

Mark Fuge
fuge@umd.edu
Department of Mechanical Engineering, University of Maryland
USA



## ABSTRACT

Interactive systems that facilitate exposure to examples can augment problem solving performance. However designers of such systems are often faced with many practical design decisions about how users will interact with examples, with little clear theoretical guidance. To understand how example interaction design choices affect whether/how people benefit from examples, we conducted an experiment where 182 participants worked on a controlled analog to an exploratory creativity task, with access to examples of varying diversity and presentation interfaces. Task performance was worse when examples were presented in a list, compared to contextualized in the exploration space or shown in a dropdown list. Example lists were associated with more fixation, whereas contextualized examples were associated with using examples to formulate a model of the problem space to guide exploration. We discuss implications of these results for a theoretical framework that maps design choices to fundamental psychological mechanisms of creative inspiration from examples.



## CCS CONCEPTS

• **Human-centered computing** → **Empirical studies in HCI**; **HCI theory, concepts and models**.

## KEYWORDS

Creativity, Examples, Interface, Problem Solving

**ACM Reference Format:**
Joel Chan, Zijian Ding, Eesh Kamrah, and Mark Fuge. 2024. Formulating or Fixating: Effects of Examples on Problem Solving Vary as a Function of Example Presentation Interface Design. In *Proceedings of the CHI Conference on Human Factors in Computing Systems (CHI '24), May 11–16, 2024, Honolulu, HI, USA.* ACM, New York, NY, USA, 16 pages. https://doi.org/10.1145/3613904.3642653





*CHI '24, May 11–16, 2024, Honolulu, HI, USA*



ACM ISBN 979-8-4007-0330-0/24/05
https://doi.org/10.1145/3613904.3642653


## 1 INTRODUCTION

Examples — descriptions or representations of possible *solutions* (or parts thereof) for the same or related problems [8, 18, 35, 83, 84] — are an integral part of the creative problem solving process. Examples can take many forms, such as previous physical prototypes brought to a brainstorm [33], search results for patents for related problems [25], spoken ideas from collaborators [66], UI designs in an accessible gallery [50]), or references from memory to earth animals or science fiction creatures when inventing new fictional alien creatures [89]). Importantly, examples can substantially shape what ideas come to mind [84, 89]. This **"structuring of imagination"** [89] is sometimes helpful "*inspiration*" that leads to more creative ideas [19, 23, 33, 80]. But examples can also have harmful "*fixation*" effects that constrain novelty and innovation [40, 54, 84, 89] (for recent reviews, see [83] and [18]). Importantly, examples can influence problem solving without conscious effort or recognition [54, 60, 61], and persist in spite of creators' explicit intentions not to be influenced by them [40, 84, 85]. Perhaps in recognition of these facts, effective creators take an active role in finding, structuring and interacting with examples [23, 35, 79] using a variety of analog and digitally mediated systems and practices, such as search engines [35, 68, 69, 97], design workbooks and commonplace notebooks [26], online whiteboards [58, 86], mood boards [55], and wider interactions with their community of practice, such as trade publications and conventions [23, 33].

An important area of HCI research on creativity support tools therefore studies the design of interactive systems that can assist creators in discovering examples [9, 39, 47, 48, 74, 80, 82, 87, 88], structuring, analyzing, and exploring collections of examples [15, 16, 53, 57, 81, 93], and adapting and using examples [39, 48, 50]. Designers of such systems need to grapple with an array of very practical interaction design decisions. For example, how should we support interaction with examples over different screen sizes? Should examples be delivered via recommendation (in small sets), a feed, or some other interaction paradigm? What information should be presented alongside an example? We would like to have a consistent theory to draw from to make these decisions. Beyond considerations of usability, we conjecture that such a theory would need to map design decisions (or classes thereof) to creativity-relevant behaviors and outcomes, ideally with a nuanced specification of

the precise benefits and costs of each design decision for these behaviors and outcomes. A theory of **human-example interaction** like this could help us design better systems with sensible defaults, prioritize and negotiate design requirements, and guide evaluation.

As a step towards developing such a theory, we conducted an experiment with 182 participants solving a controlled analog to an exploratory creativity task [7]. We systematically varied both the diversity of examples and the types of presentations: overlaid on the search environment (the "`In-Context`" condition), presented in a list (the "`List`" condition), or in a dropdown selectable menu (the "`Dropdown`" condition). The "`In-Context`" design was inspired by an emerging pattern of *contextualizing examples in the creator's workspace or problem* in HCI systems for example-based creativity [39, 47, 78, 91, 93] on the one hand, and theoretical descriptions of the use of examples to (re)formulate problems [34, 35, 43, 59, 67, 79]; the latter "`List`" and "`Dropdown`" conditions were designed to be representative of common interfaces for interacting with examples (in search results lists and pages of recommendations).

Our primary results were threefold: 1) "`List`" presentation harmed solution quality compared with "`In-Context`" or "`Dropdown`" presentation; 2) each interface condition was associated with distinct self-reported example usage strategies (notably, more usage of examples to "model" the problem space to guide exploration in the `In-Context` vs. `List` or `Dropdown` conditions, and more usage of examples to "stimulate" a specific direction of exploration in the `List` condition); and 3) the `List` condition's propensity for stimulation-based strategies was corroborated by an increased usage of "hill-climbing" strategies early on, as evidenced by analyses of `Dropdown` distance between participants' moves.

We discuss how these results, in conversation with the literature on example-based creativity support systems as well as psychological mechanisms of creativity with examples, could contribute to a theoretical framework for designing interactive systems for creative problem solving with examples.

## 2 RELATED WORK

### 2.1 Sources of empirical variability in effects of examples

Prior work has examined how the consequences of examples for creative problem solving outcomes are related to characteristics of the examples, such as their novelty [1, 6, 11, 80] (generally positive effects), conceptual distance from the problem domain [10, 19, 25, 31, 90] (mixed or curvilinear effects), and example diversity [12, 21, 29, 38, 80, 94, 96] (generally positive or contingently positive effects). Our work contributes additional empirical results on the relative contributions of (and potential interactions, in the statistical sense, between) example characteristics and example *interface* characteristics. In particular, we explore how the example characteristic of *diversity* might interact with example interface characteristics, such as whether the examples are presented as a list vs. in context of a representation of the design space. To do this, we need to also consider the cognitive mechanisms of inspiration or fixation from examples (or varying characteristics), which might be more or less afforded by example interfaces. We discuss this body of literature in the next section.

### 2.2 Theoretical insights into human-example interaction

A number of detailed in-situ studies of creators have documented a range of strategies for working with examples, ranging from simpler, more source-driven strategies like direct source adaptation [24], to more complex and reflective strategies associated with more radical transformation of examples, such as source analysis and schema-driven source selection [24], analogical reasoning [3, 27, 28, 37], and generating novel emergent features that can connect disparate attributes across examples [92]. These "processing strategies" can be described by a variety of theoretical frames from the psychological literature on creativity. We believe this theoretical level of description could facilitate our goal of synthesizing mappings between interface characteristics and effects of examples on creative problem solving outcomes. Some notable examples include basic memory mechanisms such as *priming* [63] and spreading activation [17, 73], and higher level cognitive processes such as conceptual *abstraction* and *analogical transfer* [20, 28], *conceptual combination* [92], and *problem reformulation* based on examples [22, 34, 57]. Of particular interest in our study is a contrast between priming and spreading activation mechanisms on the one hand, which are associated with lower-level conceptual influences, and problem (re)formulation, which is associated with more complex, higher-order processing of examples. In this study, we extend this literature by exploring how two specific mechanisms of processing examples — stimulation, and (re)formulation — might be helped or hindered by different example interaction interfaces. To set the context for our results, we briefly review the literature on each mechanism here.

*2.2.1 Using examples to stimulate ideation.* Spreading activation has been invoked to explain the impact of external stimuli on ideation. For instance, the "search for ideas in associative memory" (SIAM) model [66] proposes that when ideas come to mind, whether from memory, or through discussion with others or exposure to examples, they also raise the activation level of other associated concepts and features in memory, which can stimulate or inhibit ideation by making certain sets of ideas more or less likely to be generated based on the current network of associations in memory. For example, an example idea "use as paperweight" (for a design prompt to generate alternative use for a brick) may activate related concepts such as "office", or "is heavy", along with their associated concepts; subsequent ideas such as "construct a table", or "prop up a bookshelf" may then be more likely to come to mind, compared to ideas like "use as a weapon" or "makeshift goalposts for soccer". In this way, exposure to examples can shape the trajectory of ideation, and the corresponding floor or ceiling of creativity [6, 13]. In this paper, we discuss this set of mechanisms under the label "**stimulation**", to capture the intuition of examples stimulating ideation along a particular direction.

*2.2.2 Using examples to (re)formulate problems.* Past research has documented how people can use examples to construct, refine, and even reformulate their understanding of the creative problem they are trying to solve [34, 35, 43, 59, 67, 79], through processes such as intentional free-form curation of examples [56] or on mood boards [55]. For instance, Okada et al. documented how two artists used

individual artworks to shape not just their ideas, but used a process of "analogical modification" to search for and modify higher-order concepts, and their creative vision over the course of years [67]. This process of example-influenced problem formulation is related to computational and neurobiological models of cognitive search [36] which study the range of strategies that intelligent agents can use to structure their search processes: here, there is an important distinction between "model-free" search, where external feedback from the world on the agents' actions guide search in a simpler, more local, stimulus-response manner, and "model-based" search, where the agent constructs a model or representation of the task and environment (in the case of creative problem-solving, this would be the problem and/or design space [22, 30, 65]) partly on the basis of reflection its own actions and possibly observation of others' actions, and uses that model to decide where and when to explore in the task environment vs. continue to sample locally. Additionally, insight problem solving research has documented how people not only construct models, but also substantially revise them in radical ways, to solve difficult creative problems [43, 45]; this process is a key source of difficulty for creative problems, where one's initial problem formulation (e.g., key constraints or requirements), may be unhelpful [45]. In this paper, we discuss this set of mechanisms under the label "**(re)formulation**", to capture the intuition of creators leveraging examples to (re)formulate their understanding of the design space.

## 2.3 Interactive systems for interacting with examples

In this study, we are interested in understanding how the impact of examples on problem solving varies as a function of interaction design decisions for how creators will interact with examples. To set the context, we briefly review here some emerging interaction design patterns in HCI systems research into *how* (vs. when, as in recommendation systems) participants interact with sets of examples (vs. understand or modify a single example).

One higher-level design pattern involves *explorable overviews* of examples. For example, MetaMap [42] supports exploration of examples through keywords and colors and offering a playground to curate examples; RecipeScape [16] uses a map UI to present recipe examples; Sifter [70] presents large collections of image manipulation tutorials in a faceted view based on their command-level structure; the Adaptive Ideas Web tool [53] enables designers to explore and structure collections of web design examples; the Freed system [64] empowers design students to spatially organize their digital collection of examples, define relations and reflect on their interrelationships; and Cabinet [44] supports collecting and organizing of visual examples for inspiration and reference.

Another emerging design pattern can be described as *contextualizing examples in the creator's workspace or problem*, enabling designers to curate and reflect on the examples to build an understanding of their design space. For example, ReflectionSpace [78] interactively contextualizes design artifacts in project timelines (and associated comments and reflections) to promote reflection and learning; MoodCubes [39] enables designers to curate, compare, and explore suggested 3D design elements in the context of an overall 3D "cube" room layout; IdeateRelate [93] locates design examples in coordinates of similarities corresponding to the users' original ideas; the IdeaMache system provides an environment for free-form visual curation and sensemaking of creative materials in the context of a project canvas [91]; and ImageSense embeds the process of searching for, exploring, and integrating examples into both individual and shared work spaces [47].

We build on this work by directly testing how the emerging pattern of contextualizing examples might impact the effects of examples on problem solving. To facilitate downstream theoretical development, we go beyond formulations of problem solving effects and outcomes that are task-specific — such as writing code [9], designing websites [50], or designing room layouts [39] — and/or removed from creativity-specific mechanisms, such as browsing and searching and exploring, to more theoretically grounded descriptions of psychological mechanisms such as fixation and problem reformulation.

# 3 METHODS

## 3.1 The WildCat Wells Task as a Controlled Analog to Exploratory Creative Problem Solving

We experimentally investigated our research questions using a controlled analog to **exploratory creativity**, a term introduced by Margaret Boden's influential model of creativity [7] to describe a subset of creative problem solving processes that involve *exploration* within a conceptual space that is often large and complex. This conception of exploratory creative problem solving as search in a space has deep roots in research on search landscapes and innovation in organization and management science [5], as well as psychological models of problem solving [65] and creativity [71] (as reviewed in our discussion of model-based mechanisms for using examples in 2.2.2). A key insight from this literature is that local search and hill-climbing are insufficient in more rugged and complex search landscapes, because they can trap searchers in local optima; to overcome this, searchers need to find ways to explore or "jump" to new regions of the landscape [5], such as guiding search through (re)modeling of the search space [36]. This contrast between local and distant exploration is often described in terms of the shift between exploitation and exploration [5], where the latter search dynamics are more associated with innovation and creativity [71]. Note that the notion of exploratory creativity is distinct from another important class of creative processes that involve what Boden [7] calls *transformational* creativity: in this form of creativity, creators search for or construct alternative problem spaces (as discussed in the related work) [22, 43], rather than search within an existing problem space as given.

Our controlled analog is the WildCat Wells task. The name of the task takes inspiration from the real-world task of wildcat drilling[1], a form of exploratory drilling for oil and gas in an unfamiliar environment where the distribution of resource-rich locations is unknown. Accordingly, in this task, participants can "drill" for "resources" in a 2D grid by clicking on locations in the grid. Clicking on a grid location then uncovers a score amount, analogous to the amount

[1] https://en.wikipedia.org/wiki/Wildcatter

of oil/gas uncovered at a drilling site. Like its real-world counterpart, the distribution of resources in this task is unknown; in our particular instantiation, participants' goal is to uncover the most resource-rich drilling location (i.e., the grid location with the highest score). Following our conceptualization of examples as descriptions/representations of possible solutions to the same/similar problem, in this task, we operationalized examples as possible grid locations and their associated scores.

We chose this task for several reasons. First, we had a high degree of parameter control over the properties of the task and examples, which allowed us to precisely control the ruggedness of the task structure and also employ a within-subjects design while mitigating learning effects by constructing and sampling from a set of Wildcat Wells' tasks with isomorphic ruggedness/complexity properties (see Section 3.1.1). The task structure also gave us granular and precise measures of process and outcome dynamics. Finally, the simplicity of the task allowed us to minimize the impact of prior knowledge because the task does not require specialized domain expertise. While the specific task structure in terms of distribution of rewards over the search space is unknown to participants, the generic task structure of searching a space for rewards is probably not unfamiliar to most people. The Wildcat Wells task and its operationalization of examples is also conceptually similar to other instances of exploratory creativity that may draw on example solutions from a very similar problem: for instance, when searching for effective parameter settings for wing airfoil designs, other airfoil designs — which, like our grid location, are also combinations of parameters — may serve as relevant examples; when designing effective ads for a vaccine persuasion campaign, other vaccine persuasion ads — which are also combinations of design features — may serve as relevant examples; and when designing effective UI elements, other websites and their UI elements — which are also compositions of UI features — may serve as relevant examples. We also adapted the Wildcat Wells task specifically from a prior study [62] of the dynamics of exploration and exploitation in collaborative problem solving. However, because the WildCat wells task is only analogous to exploratory creativity, our results here can only speak to effects of examples on exploratory, but not transformational, creative problem solving.

*3.1.1 Search Environments.* Our WildCat Wells search environments consisted of a 100x100 grid of points (with corresponding scores controlled by a synthetic objective function that determines the distribution of scores; see Algorithm 1 in Appendix A and our source code for generating search environments). Figure 1 (A) shows a representative search environment we used in our experiment.

Our goal was to more closely match the difficult search spaces that the creativity theorist David Perkins calls "Klondike spaces" [71], which are environments where simple "hill-climbing" exploration strategies are insufficient, and likely outperformed by other creative exploration strategies such as a mix of exploration and exploitation [5, 71]. We describe the specific parameter settings and algorithm we used to generate these task environments in Appendix A (and share the code generating the environments in the Supplementary Material); here, we note that we set the parameters to yield a search environment that was fairly rugged (adding more false "peaks" to incorrectly intuit as the location of the maximum score) and locally noisy (reducing the local correlation between scores in the grid, such that searchers would often be surprised by the score of nearby regions in the grid). The parameters to generate search environments were determined by a series of rubrics (e.g. more than one area with scores higher than 80), and pilots for a qualitative sense of difficulty based on the topology of the solution space.

To reduce the likelihood that our results were tied to a specific formulation of the search environment, we generated 10 search environments with the same synthetic objective function parameters but different random seeds. The resulting search environments were qualitatively similar to Mason & Duncan's work [62].

*3.1.2 Examples.* To prepare sets of initial examples with **High Diversity** (HD) or **Low Diversity** (LD), we randomly generated 10,000 sets of 10 examples each (recall that each example is a "drilling location" point in the 100x100 Wildcat Wells search environment, with a corresponding score) for each of our 10 search environments, and ranked the diversity of each example set with a close variant of the Determinantal Point Process (DPP) approach [49] described in Algorithm 2 in Appendix B proposed by Eesh et al. [41]; intuitively, this approach measured the "hyper-volume" spanned by a selected set of points, such that larger volumes corresponded to higher levels of diversity, since these points spanned a larger set of the space of possible moves [49]. We then randomly picked three HD examples sets that were greater than the 99th percentile of the distribution of diversity across the example sets, and three LD examples sets with diversity lower than 1st percentile of the diversity distribution. To ensure that examples would not directly reveal the location of the peak, or provide a high enough score that participants might simply stop after seeing the example instead of searching, we discarded example sets that had any point with a score over 80 (scores in the search environment ranged from 0 to 100), and resampled example sets as necessary — subject to the same low/high diversity sampling criteria) to construct our final example sets for each search environment. Figure 1 (B) shows an LD and HD example set used in our experiment.

## 3.2 Experiment Design

We conducted a mixed design experiment. **Example interface** was a *between-subjects* factor, with three conditions: 1) "parallel examples with context" interface: all 10 examples were shown in the 100x100 space with color coding to denote the score associated with each point, referred to as the "`In-Context`" interface 2) "parallel examples without context" interface (shown in Figure 2): all 10 examples were shown in a list, also with color coding to indicate example score, referred to as the "`List`" interface 3) "serial examples without context" interface: only one example was shown at a time and the participant needed to use a dropdown button to see other examples, referred to as the "`Dropdown`" interface. Figure 2 shows the experimental interface of the `List` condition as an example. The `In-Context` interface was inspired by design patterns of example interfaces that contextualized examples in the creator's workspace or problem (e.g., [39, 47, 78, 91] The `List` interface was inspired by the familiar design pattern of a "list" of examples, often in the context of a search interface (as search results), or list

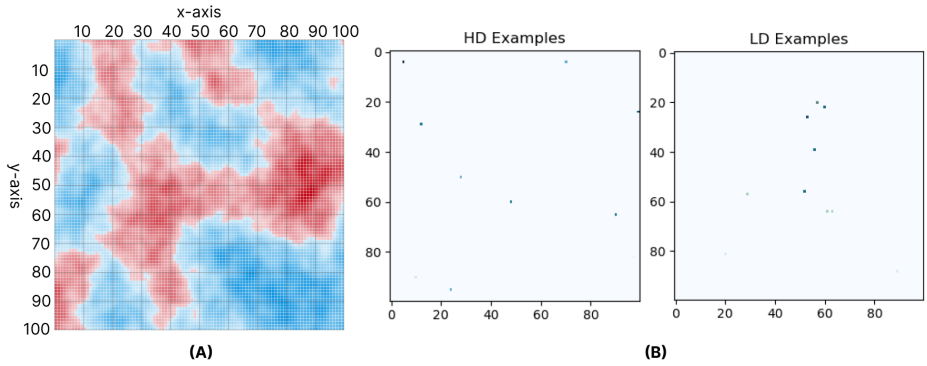



**Figure 1: Example Wildcat Wells search environment with color coding of points to indicate their scores (0-50: dark blue to light blue, 50-100: light red to dark red) (A), and example sets of high and low diversity sets of points in this search environment, which are given as examples (B).**

of recommendations in a recommender interface. The `Dropdown` interface was designed to approximate more constrained interfaces for interacting with examples, such as through chat-based or recommendation systems (e.g., popping up one or two examples at a time). The three example interfaces were shown in the context of the WildCat Wells task in Figure 3. We conjectured that interfaces that allowed for comparison between examples (whether in the context of a task environment, as in the `In-Context` interface, or just with attributes shown for comparison, as in the `List` interface) might facilitate more model-based usage of examples (what we called a "(re)modeling" mechanism in 2.2). Since we designed our Wildcat Wells task to be unsuitable for simpler hill-climbing (e.g., "stimulation-based" mechanism as described in 2.2), we also expected that these interfaces might also lead to better performance on the task, through, for example, model-based exploration strategies.

**Example diversity** was a *within-subjects* factor: each participant attempted the WildCat Wells task twice, once with a set of HD examples, and once with LD examples. Recall that we generated 10 variant search environments, each with their own set of HD and LD example sets. To approximate counterbalancing of our within-subjects factor, we created 2 "run" variants for each search environment, with each variant having an HD or LD example set as the first trial. Participants were randomly assigned first to an example interface condition, and then randomly assigned to one of the 20 potential "runs" in each interface condition (but constraining assignment such that participants would not see the same search environment twice). Based on prior research on example diversity, we expected that participants would perform better when given high vs. low diversity example sets.

## 3.3 Participants

We recruited participants from the Amazon MTurk platform, limiting participants to U.S. residents with more than 500 HITs with at least 99% approval rate. Each participant was paid US$1.3 for their participation, which was an effective rate of $10 per hour, given the average task completion time of 8 minutes.

We aimed for a total sample size of 195 (65 per each of the three conditions), to achieve target statistical power of over 0.80 to detect medium-sized statistical effects in a mixed between-within design experiment analysis. After rejecting invalid work of 42 participants for irrelevant responses (e.g., "nice") to the closing survey question about how they used examples), we obtained data from 182 participants (63 females, 118 males, 1 other; 65 in context, 56 `List`, and 61 `Dropdown`) in total, yielding an effective statistical power of 0.86 for medium-sized effects.

## 3.4 Experimental Procedures

Participants experienced the WildCat Wells task as a 100x100 space (see Figure 2). Their task was to find the square with the highest score. Participants explored squares by clicking on them to reveal their underlying score, shown in color coding, similar to the examples. To simulate the constrained nature of real creative tasks (which often have some time/budget pressure) and reduce the likelihood of ceiling effects, participants had a total budget of 60 moves for exploring squares. This budget was estimated from our pilot studies, where on average, most participants found the highest scoring square within 50 moves. We also provided incentives to encourage participants' exploration: there was a $0.25 bonus for achieving a highest score greater than 95, and a $0.50 bonus for achieving the maximum score of 100. The information panel on the right side of the experimental interface (see Figure 2 showed moves remaining, the score of the current exploration and the maximum score the participant had achieved in the current round.

Since the WildCat Wells task does not interact strongly with prior knowledge in any particular domain, we addressed potential pre-existing differences in ability by measuring participants' baseline divergent thinking ability, a correlate of creative ability [75]. Before the study, we asked participants to generate as many alternative uses of coffee cup as they could in 2 minutes (an instance of the commonly used Alternative Uses task [32] for measuring divergent thinking [75]). Participants were then given one trial round through the WildCat Wells task (without examples) to familiarize them with the interface and task. After that, participants completed two formal rounds of the WildCat Wells task, which constituted the main

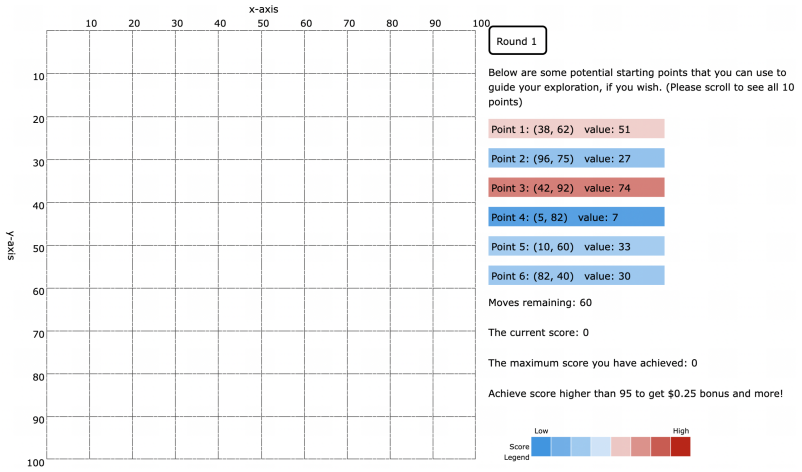



**Figure 2: Screenshot of experimental interface, shown for the `List` condition: the 100x100 grid, which constituted the search environment for the task, was shown on the left panel: participants explored the space by clicking anywhere on the 100x100 grid. The 10 initial examples, moves remaining, the score of current move, the current max score and score legend were shown on the right panel. In the `Dropdown` condition, the dropdown menu as seen in Figure 3 was shown in the same position as the list of examples in the `List` condition. In the `In-Context` condition, examples were instead overlaid as points, with corresponding values, on the search grid, as shown in Figure 3.**

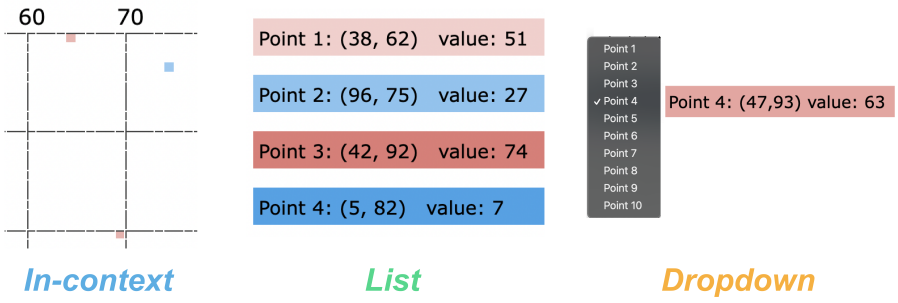



**Figure 3: Three conditions of presenting examples: "`In-Context`" (directly on the search environment grid), "`List`" (in a list) and "`Dropdown`" (in a clickable dropdown selector).**

experimental trials in our study. Finally, participants completed a post-study questionnaire, with three free-response questions: 1) What strategy did you use for hunting? 2) How did you use initial examples (the values of ten points given to you)? 3) What differences did you notice between initial examples given in those two rounds? Which did you find helpful?

We obtained institutional IRB approval for the whole project prior to the study.

## 4 RESULTS: PLANNED ANALYSES

### 4.1 No significant differences in baseline divergent thinking ability across interface conditions

We first report the results of our check for random assignment with respect to divergent thinking ability and baseline performance on our task. We observed no statistically significant difference in the number of generated alternative uses across three conditions ("`In-Context`" participants: $M = 6.52, SD = 3.02$; "`List`" participants: $M = 5.75, SD = 3.58$; "`Dropdown`" participants: $M = 6.46, SD = 3.87$, Kruskal-Wallis $H = 2.40, p = 0.30$). Similarly, we observed no statistically significant difference in participants' best score on the trial run of the Wildcat Wells task across the conditions ("`In-Context`" participants: $M = 90.97, SD = 7.52$; "`List`" participants: $M = 90.91, SD = 7.13$; "`Dropdown`" participants: $M = 91.18, SD = 7.37$, Kruskal-Wallis $H = 0.19, p = 0.91$). This suggests that participants across the interface conditions were comparable in terms of baseline divergent thinking ability as well as baseline task performance.

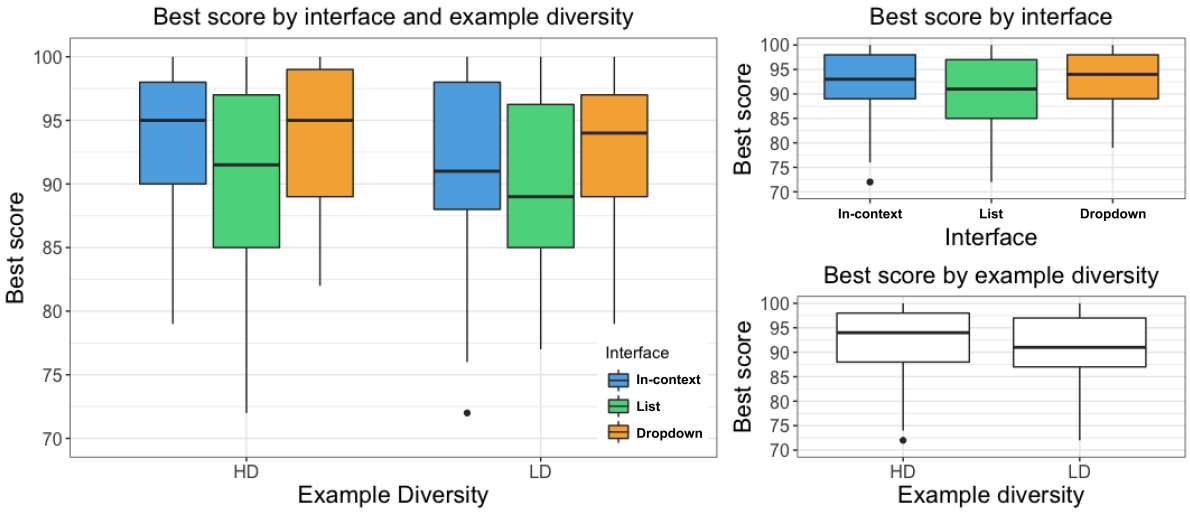



**Figure 4: Distribution of best scores by interface and example diversity conditions. Participants in the `List` interface condition had lower best scores than participants in the other interface conditions regardless of example diversity (top right). Best scores were also lower when participants were given low vs. high diversity examples (bottom right).**

### 4.2 `List` presentation of examples and low diversity example sets associated with lower best scores

The `List` condition had slightly lower scores on average compared to the other conditions (regardless of example diversity; Fig. 4, top right). There was also an overall slight advantage of HD examples over LD examples (Fig. 4, bottom right).

A linear mixed effects model with best score as the dependent variable, interface condition and example diversity as factors, and random intercepts for participants (estimated in the 'lme4' package in 'R'), showed a significant main effect of interface condition, $F(2, 179) = 5.92$, $p < .01$, $eta^2 = 0.06$. Pairwise post-hoc comparisons with Bonferroni corrections showed that participants in the `List` interface condition had significantly lower best scores (est. marginal mean = 90.4, SE = 0.65) compared to both the `In-Context` (est. marginal mean = 92.7, SE = 0.61, contrast t ratio = 2.54, $p < .05$) and `Dropdown` interface conditions (est. marginal mean = 93.4, SE = 0.63, contrast t ratio = 3.31, $p < .01$).

There was also a significant main effect of example diversity, $F(1, 181) = 4.59$, $p < .05$, $eta^2 = 0.02$; post-hoc comparisons with Bonferroni corrections showed that participants had higher best scores when they received HD (est. marginal mean = 92.7, SE = 0.46) vs. LD examples (est. marginal mean = 91.5, SE = 0.46, contrast t ratio = 2.14, $p < .05$).

## 5 RESULTS: EXPLORATORY ANALYSES

We conducted a set of exploratory analyses to better understand the results of our main planned analyses, focusing on understanding process effects of interface conditions that might plausibly explain performance differences.

### 5.1 `In-Context` associated with early performance advantages, and `List` with early and persistent performance disadvantages

First, for a more granular view of performance, we examined how the participants' best score changed as a function of their move sequence. This analysis confirmed a cumulative disadvantage for participants in the `List` condition, but also showed an early advantage for the `In-Context` interface, particularly with LD examples (see Figure 5). Using a Kruskal-Wallis H-test on the current max score from the 1st to the 30th move, we observed statistically significant differences from the 1st move to the 6th move and the 8th move except the 7th move (see Table 1).

### 5.2 Variations in example presentation interfaces associated with different self-reported example usage strategies

Next, to understand how participants used the initial examples in their exploration, two researchers coded participants' responses to the question "How did you use initial examples (the values of ten points given to you)?" with three codes: **not using**, **stimulation-based** and **model-based**. This classification was guided and refined by our initial theoretical interest in the contrast between stimulation-based and (re)modeling-based use of examples, as discussed in 2.2. Examples of responses coded as "not using" include *"I did not give much thought to it"*, and *"Not much to be honest"*; examples of "stimulation-based" responses included *"Start at the reddest one and explore its surroundings"*, and *"I looked around the higher values for boxes that were darker"*; examples of "model-based" responses include *"To get an overview on which squares would be best"*, and *"They gave a vague idea of whether or not there might*

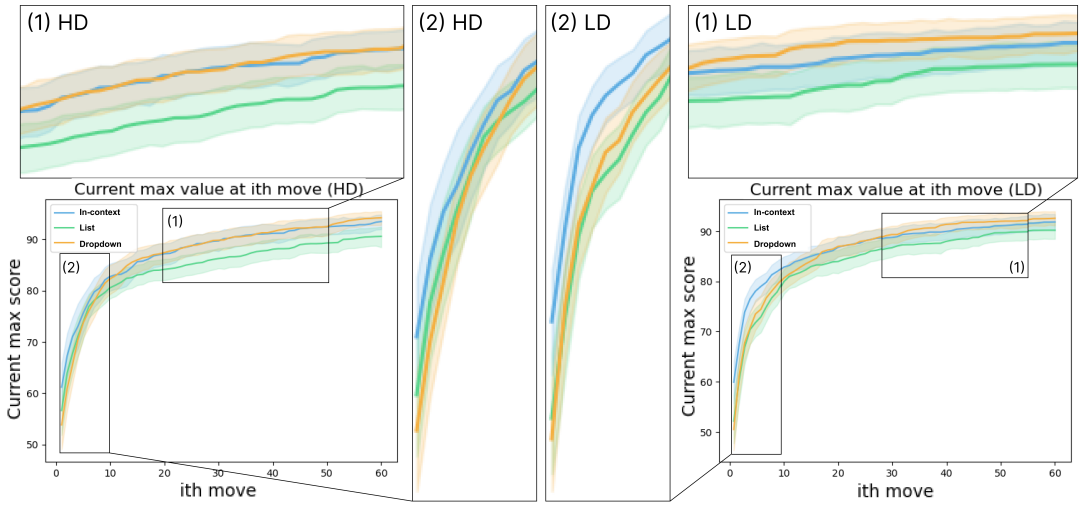



**Figure 5: Maximum score at n-th move for each participant: left) HD; right) LD. We observe (1) cumulative disadvantages for the `List` condition, as well as (2) early advantages for the `In-Context` interface, especially with LD examples.**

| ***n*-th move** | **`In-Context` (*M*)** | **`List` (*M*)** | **`Dropdown` (*M*)** | **Statistic** | **p-value** |
|---|---|---|---|---|---|
| 1 | 60.06 | 52.30 | 50.64 | 9.13 | 0.01 |
| 2 | 67.65 | 60.45 | 61.39 | 6.93 | 0.03 |
| 3 | 74.00 | 66.95 | 67.89 | 9.58 | 0.008 |
| 4 | 76.65 | 70.54 | 70.84 | 9.06 | 0.01 |
| 5 | 78.20 | 71.93 | 73.66 | 10.13 | 0.006 |
| 6 | 79.15 | 73.07 | 74.59 | 9.59 | 0.008 |
| 7 | 79.98 | 75.07 | 76.87 | 5.93 | 0.06 |
| 8 | 81.46 | 76.86 | 78.21 | 6.91 | 0.03 |

**Table 1: Means (*M*) and results of Kruskal-Wallis H-test on the current best score for three interface conditions with LD examples. There were statistically significant differences between the conditions from the 1st to 8th moves (p<0.05 except the 7th move).**

*be "hot" or "cool" zones around those points"*. When we could not infer how the participants used the initial examples, the answers were coded as "unclear". The researchers were blinded to condition during coding. Inter-rater reliability was substantial, at Cohen's $\kappa$ = 0.725 [51]; all disagreements were resolved by discussion.

The `In-Context` condition had the largest portion (30.8%) of participants who self-reported using the initial examples to *model* the space, compared to the `List` condition (10.7%) and the `Dropdown` condition (11.5%) (see Figure 6). In contrast, for self-reported use of initial examples to *stimulate* their exploration, the `List` condition had the highest percentage (42.9%) followed by the `Dropdown` condition (29.5%) and the in context condition (29.2%). Finally, 17/61 (27.9%) `Dropdown` participants self-reported that they were not using examples, which was higher than participants in the other two conditions. Our log data were consistent with this observation: 37/61 (60.7%) participants in the `Dropdown` condition never clicked the dropdown button to see other examples in both HD and LD conditions. Of the remaining participants who did click on the dropdown button, we infer — assuming that each subsequent click corresponds to an example view — that the mean number of examples viewed was M = 7.37 (SD = 3.40) for HD, and M = 7.05 (SD = 3.61) for LD.

*5.2.1* `In-Context` *participants more likely than other interface conditions to self-report model-based example usage, and* `Dropdown` *participants more likely than other conditions to self-report not-using examples.* We first statistically tested these patterns with a series of logistic regressions, one for each example strategy (not-using, stimulation-based, and model-based) (see Table 2). We ran separate logistic regressions rather than a single multinomial regression given our interest at this step in the relative likelihood across interface conditions of self-reporting a particular example strategy, rather than relative differences across strategies within each condition (best answered by a multinomial logistic regression). We first observe that participants in the `List` and `Dropdown` conditions were less likely to self-report using a model-based strategy compared to the `In-Context` condition ($B$ = -1.31, 95% CI = [-2.31, -0.31], z = -2.57, $p < .05$ for `List` vs. `In-Context`, and $B$ = -1.23, 95% CI = [-2.18, -0.29], z = -2.55, $p < .01$ for `Dropdown` vs. `In-Context`). In more intuitive terms, `In-Context` participants were approximately 3x more likely to self-report using a model-based example usage strategy

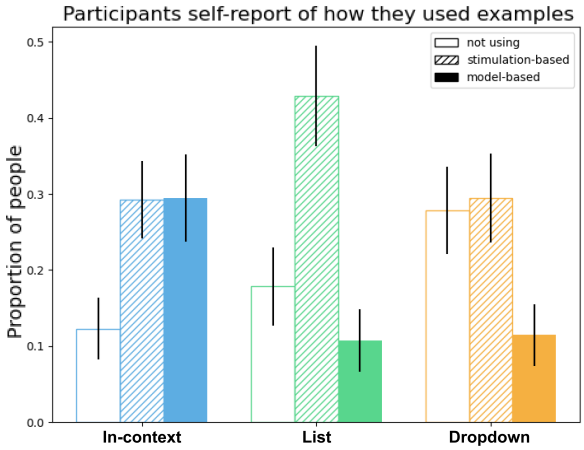



**Figure 6: Raw proportion of participants expressed "not using (examples)", "stimulation-based" or "model-based" in their answer to "How did you use initial examples (the values of ten points given to you)?". Error bars are standard error of proportion. More participants self-reported using a model-based strategy in the `In-Context` condition compared to other conditions.**

compared to `List` or `Dropdown` participants (Odds Ratio = 3.7 and 3.4 for `In-Context` vs. `List`, and `In-Context` vs. `Dropdown`). The overall model fit was better than a null model ($LL_{model}$ = -80.93 vs. $LL_{null}$ = -86.16), Likelihood Ratio $\chi^2(2) = 5.23$, $p < .01$. Next, we observe that participants in the `In-Context` condition were less likely to self-report not using examples compared to the `Dropdown` condition ($B$ = -1.01, 95% CI = [-1.94, -0.09], z = 2.14, $p < .05$); in more intuitive odds ratio terms, `Dropdown` participants were 2.7x more likely than `In-Context` participants to self-report a "not using" strategy (Odds Ratio = 2.75). The overall model fit, though better than a null model ($LL_{model}$ = -86.62 vs. $LL_{null}$ = -89.10), was marginally significant, Likelihood Ratio $\chi^2(2) = 5.14$, $p = .08$. Finally, we observe that there were no significant differences across conditions in the likelihood of self-reporting a stimulation-based strategy. Indeed, the overall model fit, though nominally better than the null model ($LL_{model} = -114.52$ vs. $LL_{null} = -116.08$ was not statistically significant, Likelihood Ratio $\chi^2(2) = 1.56$, $p = .21$.

*5.2.2* `List` *participants more likely to self-report a stimulation-based example usage strategy compared to not-using or model-based example usage.* Next, we focus on statistically evaluating the apparent predominance of a stimulation-based self-reported example usage strategy for `List` participants. We fitted a multinomial logistic regression, with model-based usage as the reference outcome class. Participants in the `List` condition were significantly more likely to self-report using a stimulation-based strategy compared to a model-based one, $B = -1.44$, 95% CI = [-0.34, -2.53], z = -2.58, $p < .01$. In odds ratio terms, participants in the `List` interface condition were 4x more likely to self-report a stimulation-based vs. model-based strategy (Odds Ratio = 4.21). The overall model fit was statistically significantly better than a null model, Likelihood Ratio $\chi^2(4) = 14.39$, $p < .01$.

## 5.3 `List` presentation of examples associated with more local initial exploration of the solution space

Finally, we explored how log data might be consistent (or not) with participants' self-reported example usage strategies. We wanted to study how initial examples would affect participants' exploration behaviors, especially at the beginning of exploration when the examples provided were a major source of information. To explore this, we first constructed an exploration graph for the first 30 moves of each participant trial by computing Euclidean distances between each successive move; the intuition was that long sequences of low distances between moves would suggest "hill-climbing", and large distances would suggest "jumps". We conjectured that a "hill-climbing"-like exploration graph would be consistent with a stimulation-based strategy, rather than a model-based strategy.

Two coders independently coded all 364 exploration graphs (each of the 182 participants had an HD plot and a LD plot), coding whether the exploration behaviors were hillclimbing (h; relatively low and flat exploration graphs, with no large jumps) or not (n) for sequences of 10 moves (the 0th-10th, the 10th-20th, the 20-30th). Two examples of coding are shown in Figure 7. We coded 1092 10-move instances (3 10-move instances per round x 2 rounds per participant x 182 participants) with substantial inter-rater reliability, Cohen's $\kappa$ = 0.78.

For all trials, a higher proportion of participants in the `List` interface condition used an exclusively hill-climbing strategy for the first 30 moves (proportion=0.161, lower bound=0.112, upper bound=0.210) compared to the `In-Context` (proportion=0.046, lower bound=0.020, upper bound=0.072) and `Dropdown` condition (proportion=0.082, lower bound=0.047, upper bound=0.117; see Figure 8aa). Similarly, for LD trials, the proportion of participants in using an exclusively hill-climbing strategy for the first 30 moves was higher for the `List` condition (34%) compared to the `In-Context` (18%) and `Dropdown` condition (18%) (Fig. 8ab, left). In contrast, for HD trials, the proportion of participants using an exclusively hill-climbing strategy for the first 30 moves was similar for the `In-Context` (18%), `List` (25%), and `Dropdown` conditions (23%) (Fig. 8ab, left).

To statistically test these observations, we fitted a series of logistic regressions, estimated with maximum likelihood, predicting $p(all\_hill)$, the probability of being all hill-climbing in the first 30 moves as a function of $interface$. Prior work suggests that choice of exploration vs. exploitation is influenced by the "goodness" of the current region of the search space (better scores makes hill-climbing more likely) [5, 36]. Our data confirmed this pattern: the average score of the first move in each of the first three 10-move blocks was positively correlated with the likelihood of being all hill-climbing in the first 30 moves, Kendall's $\tau$ = .17, $p < .01$. Thus, we conditioned our logistic regression models on the average score at the beginning of each 10-move block.

We first analyzed $p(all\_hill)$ aggregated across both HD and LD trials (value would be 1 if both LD and HD trials were 1), and then HD and LD trials separately. Table 3 shows the coefficient estimates for each of these models. For **all trials**, the coefficient for the contrast between the `List` and `In-Context` conditions was $B$ = -1.79, 95% CI=[-3.40, -0.45], z = -2.45, $p < .05$; in odds ratio terms,

| | Not using | Stimulation-based | Model-based |
|---|---|---|---|
| Intercept | -0.951 [-1.511, -0.391] | -0.288 [-0.817, 0.242] | -0.811 [-1.338, -0.284] |
| `In-Context` | **-1.013** [-1.940, -0.085]* | -0.597 [-1.349, 0.156] | |
| `List` | -0.575 [-1.459, 0.309] | | **-1.309** [-2.307, -0.312]** |
| `Dropdown` | | -0.583 [-1.347, 0.180] | **-1.232** [-2.179, -0.285]* |

**Table 2: Coefficient estimates from multinomial logistic regressions of probability of self-reported example usage strategy (1 each for not-using, stimulation-based, model-based) on interface condition. Statistics reported as *"coefficient, 95% CI ([lower, upper])"*. When the cell for a given interface condition is blank, that condition was used as the reference class in the regression. *p < .05; **p < .01; ***p < .001.**

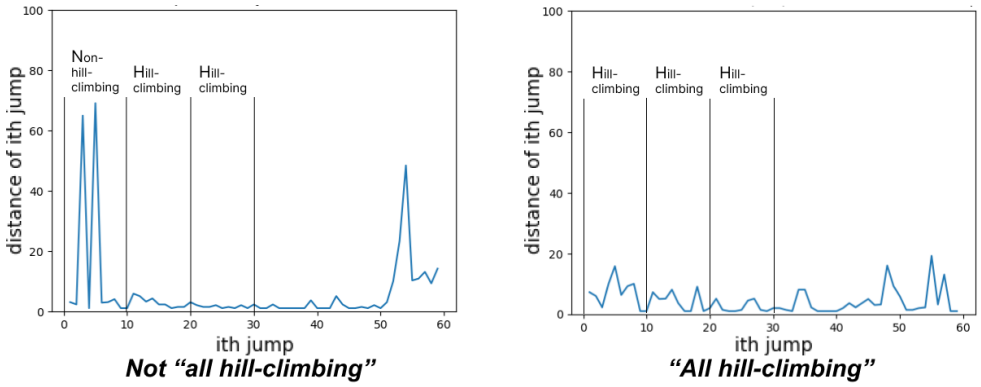


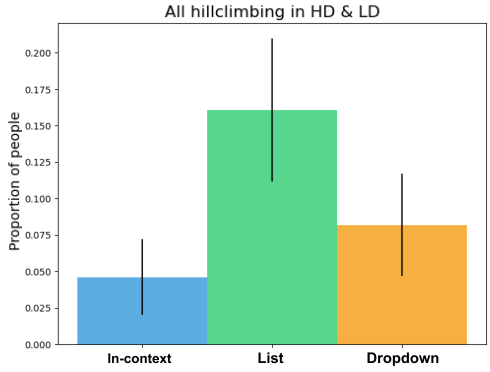

**Figure 7: Example exploration graphs used for coding hill-climbing strategies: not "all hill-climbing" (left) and "all hill-climbing" (right), where each transition between participant moves is plotted on the x-axis, and the Euclidean distance between each move and its immediately preceding move is plotted on the y-axis. In the not "all hill-climbing" example, the first 10 moves (the 0th-10th moves), where there are substantial variations in `Dropdown` move distances across the sequences, would be coded as non-hillclimbing (N), while the 10th-20th moves and the 20th-30th moves, where `Dropdown` move distances are consistently low, would be coded as hillclimbing (H). In the "all hill-climbing" example, all first 30 moves would be coded as hillclimbing (H).**

participants in the `List` condition were 6x more likely to use an exclusively hill-climbing strategy for the first 30 moves, compared to participants in the `In-Context` condition (Odds Ratio = 5.99). Note that this effect was independent of the significant positive coefficient for the average first score in the block. The overall model fit was statistically significantly better than a null model ($LL_{model}$ = -48.42 vs. $LL_{null}$ = -56.48), Likelihood Ratio test $\chi^2(3) = 16.12$, $p < .01$. Similarly, for **LD trials**, there was a statistically significant coefficient in the logistic regression model for the contrast between the `In-Context` and `List` conditions, $B$ = -0.95, 95% CI=[-1.83, -0.10], z = -2.16, $p < .05$. In odds ratio terms, participants in the `List` condition were 2.5x more likely to use an exclusively hill-climbing strategy for the first 30 moves, compared to participants in the `In-Context` condition (Odds Ratio = 2.58). There was a similar contrast between the `Dropdown` and `List` conditions, $B$ = -0.95, 95% CI=[-1.85, -0.10], z = -2.12, $p < .05$. As with the all trials model, this effect was independent of the significant positive coefficient for the average first score in the block. The overall model fit was statistically significantly better than a null model ($LL_{model}$ = 92.98 vs. $LL_{null} = -98.32$), Likelihood Ratio test $\chi^2(3) = 10.67$, $p < .05$. In contrast, there were no statistically significant contrasts between conditions in the **HD trials**, though the numerical pattern of results were similar to the other models (generally negative coefficients for `In-Context` and `Dropdown` vs. `List` conditions). The overall model fit, though nominally better than a null model ($LL_{model} = -95.14$ vs. $LL_{null} = -95.85$, was not statistically significant, Likelihood Ratio $\chi^2(3) = 1.42$, $p = .70$.

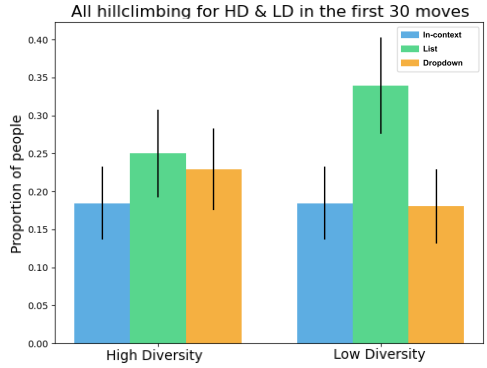
Because we were concerned this pattern of differences might be driven by pre-existing individual differences in propensity to hill-climbing, rather than a shift due to the interface condition, we repeated our coding procedure for exploration graphs generated from the initial trial round of exploration, which did not include examples (described in 3.4. The proportion of participants who displayed a predominant hill-climbing strategy, as described above, was distributed across conditions as follows: `In-Context`= 0.05 (SE = .03), `List`= .07 (SE = .03), and `Dropdown`= 0.15 (SE = .05). A logistic regression predicting the probability of a predominant hill-climbing strategy (yes or no) as a function of interface did not improve fit over a null model with no predictors, $\chi^2$ (2) = 4.17, $p = .12$; note, however that the overall frequency of hill-climbing strategies was lower than in the main trials, and the `List` condition was not the condition with highest frequency (in contrast to the main trials). There was also no significant correlation between the likelihood of hill-climbing predominance and either the number of alternative uses task responses, r = .04, $p = .61$, or the likelihood of hill-climbing predominance in the main trials, r = .03, $p = .65$. Altogether, these

| | All trials | HD trials | LD trials |
|---|---|---|---|
| Intercept | -7.322 [-11.73, -3.739] | -1.734 [-3.522, -0.052] | -2.551 [-4.409, -0.870] |
| `In-Context` vs. `List` | **-1.789** [-3.400, -0.449]* | -0.431 [-1.325, 0.448] | **-0.950** [-1.833, -0.103]* |
| `Dropdown` vs. `List` | -0.926 [-2.220, 0.260] | -0.113 [-1.325, 0.448] | **-0.946** [-1.850, -0.0858]** |
| Avg. first score in block | **0.089** [ 0.035, 0.152]** | 0.010 [-0.016, 0.037] | **0.031** [ 0.005, 0.059]* |

**Table 3: Coefficient estimates from logistic regressions of probability of predominantly hill-climbing strategy in the first 30 moves on interface condition and average first score in block, across all, HD, and LD trials. Statistics reported as *"coefficient, 95% CI ([lower, upper])"*. *p < .05; **p < .01; ***p < .001.**




**(a)**




**(b)**

**Figure 8: Raw proportion of participants with a predominantly hillclimbing strategy in the first 30 moves, across interface conditions (a) with both HD and LD examples (b). More "`List`" participants did hillclimbing in the first 30 moves with both high and low diverse example sets than "`In-Context`" participants. More "`List`" participants did hillclimbing for the first 30 moves with LD examples than other combinations of the presentation and example sets.**

results are inconsistent with the alternative explanation that `List` participants were simply more likely (due to individual differences) to choose a predominantly hill-climbing strategy overall; instead, taken together with the survey results, we believe this set of results suggest a shift in strategy towards hill-climbing (or, as we describe in this paper, a stimulation-based strategy for using examples).

# 6 DISCUSSION AND CONCLUSION

## 6.1 Summary and Interpretation of Results

In this paper, we aimed to contribute to a theory of human-example interaction to guide the design of example-based creativity support tools. Towards this goal, we conducted an experiment with a controlled analog to an exploratory creativity [7] task to investigate how example presentation interface variations influence whether/how people benefit from examples.

We found evidence that **`List` presentation of examples might harm the quality of final solutions**. For example, we found variations across interface conditions in mean best score obtained at the end of trials across example interface and diversity conditions (`List` participants had worse best scores compared to `In-Context` and `Dropdown` participants; Section 4.2) and cumulative performance differences (with an early and persistent disadvantage of the `List` condition compared to the other conditions, with an especially pronounced early disadvantage relative to the `In-Context` condition; Section 5.1). This result is conceptually significant because the `List` participants received more information (seeing all 10 examples at the same time) than the `Dropdown` participants (only seeing 1 example at a time. and over 60% of them never checked other examples), and approximately equivalent information but different presentation compared with the `In-Context` condition. This suggests that seemingly unimportant, low-level interaction design decisions with respect to presentation of examples can have measurable consequences for creative problem solving performance. Separately, we also observed **beneficial effects of diversity for final solution quality**, in line with some previous work [2, 4, 29, 38, 80, 96]; importantly, this effect was similar in magnitude to the example presentation effects, suggesting that example presentation considerations may be just as important to consider as example characteristics when designing example-based creativity support systems.

Second, our exploratory analyses suggest that **`In-Context` and `List` presentation of examples may lead to distinct patterns of example usage**. `In-Context` presentation of initial examples was associated with a greater likelihood of a "model-based" strategy for using examples, where participants self-reported using the examples to gain an overall understanding of the distribution of score in the search environment to guide their exploration, compared to `List` or `Dropdown` presentations. Conversely, the `List` presentation of initial examples seemed to encourage a predominant "stimulation-based" example usage strategy, where participants selected promising examples as starting points for their exploration. Importantly this self-report data was consistent with patterns in our log data: we observed that `List` participants were more likely

to use a predominantly "hill-climbing" strategy (with low Dropdown distance between their moves) early in their exploration, relative to the In-Context and Dropdown participants; this association was independent of the relationship between hill-climbing behavior and the "goodness" of initial moves (hill-climbing in a given block of moves was more likely when the initial move was higher-scoring, consistent with prior empirical work on exploration/exploitation decisions [5, 36]). Considering these results alongside the performance results suggests that List participants were being *fixated* [40] by the examples.

A fruitful direction for further research would be to investigate the mechanisms that drive fixation in the List condition. One reason might be the upper limit in scores (no more than 80/100) on the examples presented to the participants; if taken as starting points to begin hill-climbing, those relatively low quality examples could be misleading, and block access to high-quality solutions. Another reason might be the increased effort needed to connect examples of the List condition to the search space, which would be consistent with past research on the cognitive load benefits of integrating diagrams and text (similar to integrating examples and the search space) in instructional design [14]. We view the difficulty of transferring from the text modality of lists of examples to the visuospatial modality of the In-Context solution space as a potential *mechanism* by which example presentation variations might shape their impact on ideation: future work could investigate in more detail how different example presentation designs might shift the *cost structure* of different processing strategies, in a similar way that variations in environment or interface structure have been shown to shape sensemaking by changing the cost structure of various crucial actions, such as skimming/previewing, moving documents, applying schemas to documents, or adjusting schemas [72, 76, 77].

Separately, we observed that the Dropdown presentation was associated with limited usage of the examples: many Dropdown participants self-reported not using examples (moreso than In-Context participants, for example), and this was also corroborated in their log data (via a lack of interaction with the example interface). We do not think that this lack of example usage is indicative of a lack of engagement: recall, for instance, that performance in this condition was on par with the In-Context condition (i.e., higher than in the List condition). Post-survey comments indicating enjoyment and engagement (e.g., "Fun game. Thank you!") were also seen across conditions at similar rates, and there were no statistically significant mean differences across the conditions in the trial run of the task. For these reasons, we believe that — possibly due to the interaction affordances — the Dropdown condition appeared to act similarly to a "no-examples" control condition, where participants used a wider mix of strategies vs. a particular set of example-based strategies tied to an experimental intervention. In light of this, the overall strong performance of the Dropdown condition is akin to past observations of strong performance by control "no-intervention" conditions in ideation experiments (see, e.g., [11, 80]; we thus add to a growing body of evidence that it may be easier to harm rather than help creative ideation by intervening (as in the List condition).

Overall, our results suggest that interaction design considerations for human-example interaction go beyond usability: there is indeed a space of mappings to explore between design affordances and fundamental psychological mechanisms of creative inspiration from examples. From a practical standpoint, our empirical results suggest the limitations of only showing examples without the problem space as context, especially if the problem space is large (a common feature of real world problems) and there exist some potential solutions far away from the initial examples. This implication is significant since the List view of examples - examples presented in a list - is commonly used in current creativity support tools, such as search engines and recommendation systems, yet was associated with substantial negative effects on the usage of examples and task performance relative to In-Context presentation of examples.

## 6.2 Limitations

The WildCat Wells task we used in our experiment is simpler than most real-world exploratory creativity tasks— such as airfoil design, ad design, or UI design — of which it is an analog. For instance, the task did not require any specialized domain knowledge, and the generic task structure of searching a space for rewards is probably familiar to most people: indeed, one participant in our study noted that in the post-survey that the task was "very fun and somewhat similar to minesweeper." Additionally, although we carefully constructed our Wildcat Wells task surfaces to be rugged, with multiple peaks of good solutions, our task technically has a single best solution; in contrast, many real-world creative problems — such as policy design — lack a single best solution, due to task factors such as intrinsic tradeoffs between different problem requirements; in these cases, creators often search for and construct "good enough" solutions under high uncertainty (though this might sometimes be a function of feasibility constraints rather than intrinsic properties of the task). It is unsurprising, then, that participants performed relatively well as a whole, and Dropdown participants also had competitive performance even though they did not interact or use the examples. We note, however, that performance was not quite at ceiling: only 16% of participants reached the global max in either trial (and 42% reached the threshold score of 95 for the first bonus. Still, caution is warranted when generalizing to other more complex instances of exploratory creativity; for instance, it may be that the effects of examples, and the corresponding effects of variations in their presentation interactions, will become more pronounced in more sophisticated tasks.

Relatedly, the Wildcat Wells task captures aspects of search dynamics (exploration and exploitation) in exploratory creative problem solving quite well, but does not enable observation of more sophisticated psychological mechanisms for working with examples. For instance, it is unclear what it might mean to "combine" different problem solving moves for this task. Additionally, while participants engaged in modeling of the problem space, they were not able to make larger changes to the problem space, such as questioning assumptions or relaxing constraints [45], or even changing the goal/problem altogether [43], mechanisms that are common in real-world creative problem solving tasks, such as design [22]. Thus, we reiterate that our results cannot speak to how example presentation design decisions might influence example usage for transformational creativity [7] tasks.

Thus, more work is needed to extend our exploration of patterns in example interaction design choices to more complex settings: for example, what might it mean to design a "contextualized" presentation of examples for UI elements, more complex airfoil designs, ad persuasion campaigns, research papers, or policy ideas? We are keen to build on existing design patterns similar to this in previous systems such as ReflectionSpace [78], MoodCubes [39], and ImageSense [47], as discussed in Section 2.3. Our implementation of the `Dropdown` condition may also be quite different from other `Dropdown` presentations, such as forward/backward interfaces (e.g., image suggestions [46]). Future studies can explore the consequences of these differences. For now, we note that our main results on the contrast between `In-Context` and `List` conditions are independent of this limitation, and recommend caution in generalizing the results from the `Dropdown` condition around non-use of examples.

Finally, we did not measure demographic information that may have been correlated with task performance or example usage and/or exploration patterns – for example, personality traits such as disagreeableness or extraversion may be correlated with real-world creative achievement [95]; and gender might interact with potential differences in visuospatial reasoning demands between example interfaces, given some existing research on gender differences in spatial ability [52].

## 6.3 Towards an interaction-oriented theory of creative inspiration from examples

Returning to our higher-level goal of constructing an interaction-oriented theory of human-example interaction, we now reflect on how the empirical results from our study, in conversation with theoretical mechanisms and design patterns from prior work, could contribute to an overall theory that bridges design patterns to psychological mechanisms.

We conjecture that a useful theory of human-example interaction could be conceptualized as paths through multiple coordinated spaces of **example interaction patterns**, **example-ideation psychological mechanisms**, **ideation characteristics** and **creative outcomes**. Paths through this overall set of coordinated spaces could then represent a set of principled design hypotheses about how to best support creative work with examples. For instance, bringing our empirical results in conversation with the literature we reviewed in sections 2.2 and 2.3, we could hypothesize that, given a particular exploratory creativity task environment like our instantiation of the WildCat wells task, where the key *creative outcome* of **solution quality** is determined at least in part by the *ideation characteristic* of **diversity of search**, which is in turn positively influenced by the *psychological mechanism* of **(re)modeling**, and negatively influenced by the mechanism of **stimulation**, it may be advantageous to choose *example interaction patterns* like **contextualizing examples in the problem space** (which is positively mapped to (re)modeling mechanisms), over patterns like **`List` viewing of examples** (which is positively mapped to stimulation mechanisms). Multiple other hypothesized paths could be generated and refined to map other example interaction patterns from prior work, such as faceted search systems, or example dissection/analysis, to other psychological mechanisms, such as conceptual combination, or analogical abstraction; each of these mechanisms might then in turn be contextually important for certain kinds of creative problems, such as policymaking or room layout design.

We believe that fleshing out these paths through these coordinated spaces towards a theory of human-example interaction can both make contributions to fundamental HCI theory — by enhancing synthesis of design knowledge about how to best support creative inspiration from examples — and practice — by providing a principled framework that is sufficiently granular and directly connected to design decisions, to guide effective design decisions when building example-based creativity support systems, and to practicing creators who wish to more effectively leverage examples in their creative process. We invite the rest of the creativity support systems community to join us in these efforts.

## ACKNOWLEDGMENTS

This material is based upon work supported by the National Science Foundation under Grant No. 1826083.

## REFERENCES

[1] Marine Agogué, Akin Kazakçi, Armand Hatchuel, Pascal Le Masson, Benoit Weil, Nicolas Poirel, and Mathieu Cassotti. 2014. The Impact of Type of Examples on Originality: Explaining Fixation and Stimulation Effects. *The Journal of Creative Behavior* 48, 1 (2014), 1–12. https://doi.org/10.1002/jocb.37 00000 _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/jocb.37.

[2] Niek Althuizen and Berend Wierenga. 2014. Supporting Creative Problem Solving with a Case-Based Reasoning System. *Journal of Management Information Systems* 31, 1 (2014), 309–340. https://doi.org/10.2753/MIS0742-1222310112

[3] L. J. Ball, T. C. Ormerod, and N. J. Morley. 2004. Spontaneous analogising in engineering design: a comparative analysis of experts and novices. *Design Studies* 25, 5 (2004), 495–508.

[4] Jonali Baruah and Paul B. Paulus. 2011. Category assignment and relatedness in the group ideation process. *Journal of Experimental Social Psychology* 47, 6 (2011), 1070–1077. 00000.

[5] Oliver Baumann, Jens Schmidt, and Nils Stieglitz. 2019. Effective Search in Rugged Performance Landscapes: A Review and Outlook. *Journal of Management* 45, 1 (Jan. 2019), 285–318. https://doi.org/10.1177/0149206318808594 Publisher: SAGE Publications Inc.

[6] Justin M. Berg. 2014. The primal mark: How the beginning shapes the end in the development of creative ideas. *Organizational Behavior and Human Decision Processes* 125, 1 (2014), 1–17. https://doi.org/10.1016/j.obhdp.2014.06.001 00056.

[7] M. A. Boden. 2004. *The creative mind: Myths and mechanisms*. New York. 00000.

[8] Joel Brandt, Mira Dontcheva, Marcos Weskamp, and Scott R. Klemmer. 2010. Example-centric Programming: Integrating Web Search into the Development Environment. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '10)*. ACM, New York, NY, USA, 513–522. https://doi.org/10.1145/1753326.1753402

[9] Joel Brandt, Mira Dontcheva, Marcos Weskamp, and Scott R. Klemmer. 2010. Example-centric Programming: Integrating Web Search into the Development Environment. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '10)*. ACM, New York, NY, USA, 513–522. https://doi.org/10.1145/1753326.1753402

[10] Joel Chan, Steven P. Dow, and Christian D. Schunn. 2015. Do The Best Design Ideas (Really) Come From Conceptually Distant Sources Of Inspiration? *Design Studies* 36 (2015), 31–58. https://doi.org/10.1016/j.destud.2014.08.001 00105.

[11] Joel Chan, Katherine Fu, Christian. D. Schunn, Jonathan Cagan, Kristin L. Wood, and Kenneth Kotovsky. 2011. On the benefits and pitfalls of analogies for innovative design: Ideation performance based on analogical distance, commonness, and modality of examples. *Journal of Mechanical Design* 133 (2011), 081004. https://doi.org/10.1115/1.4004396 00304.

[12] Joel Chan and Christian D. Schunn. 2015. The importance of iteration in creative conceptual combination. *Cognition* 145 (Dec. 2015), 104–115. https://doi.org/10.1016/j.cognition.2015.08.008

[13] Joel Chan, Pao Siangliulue, Denisa Qori McDonald, Ruixue Liu, Reza Moradinezhad, Safa Aman, Erin T. Solovey, Krzysztof Z. Gajos, and Steven P. Dow. 2017. Semantically Far Inspirations Considered Harmful?: Accounting for

Cognitive States in Collaborative Ideation. In *Proceedings of the 2017 ACM SIGCHI Conference on Creativity and Cognition (C&C '17)*. ACM, New York, NY, USA, 93–105. https://doi.org/10.1145/3059454.3059455

[14] Paul Chandler and John Sweller. 1992. The Split-Attention Effect as a Factor in the Design of Instruction. *British Journal of Educational Psychology* 62, 2 (1992), 233–246. https://doi.org/10.1111/j.2044-8279.1992.tb01017.x _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.2044-8279.1992.tb01017.x.

[15] Kerry Shih-Ping Chang and Brad A. Myers. 2012. WebCrystal: understanding and reusing examples in web authoring. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, Austin Texas USA, 3205–3214. https://doi.org/10.1145/2207676.2208740

[16] Minsuk Chang, Leonore V. Guillain, Hyeungshik Jung, Vivian M. Hare, Juho Kim, and Maneesh Agrawala. 2018. RecipeScape: An Interactive Tool for Analyzing Cooking Instructions at Scale. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. ACM, Montreal QC Canada, 1–12. https://doi.org/10.1145/3173574.3174025

[17] Allan M. Collins and Elizabeth F. Loftus. 1975. A spreading-activation theory of semantic processing. *Psychological Review* 82, 6 (1975), 407–428.

[18] Nathan Crilly and Carlos Cardoso. 2017. Where next for research on fixation, inspiration and creativity in design? *Design Studies* 50 (May 2017), 1–38. https://doi.org/10.1016/j.destud.2017.02.001

[19] Darren W. Dahl and Page Moreau. 2002. The Influence and value of analogical thinking during new product ideation. *Journal of Marketing Research* 39, 1 (2002), 47–60. https://doi.org/10.1509/jmkr.39.1.47.18930 00714.

[20] Zijian Ding, Arvind Srinivasan, Stephen Macneil, and Joel Chan. 2023. Fluid Transformers and Creative Analogies: Exploring Large Language Models' Capacity for Augmenting Cross-Domain Analogical Creativity. In *Proceedings of the 15th Conference on Creativity and Cognition (C&C '23)*. Association for Computing Machinery, New York, NY, USA, 489–505. https://doi.org/10.1145/3591196.3593516

[21] Alex Doboli, Anurag Umbarkar, Varun Subramanian, and Simona Doboli. 2014. Two experimental studies on creative concept combinations in modular design of electronic embedded systems. *Design Studies* 35, 1 (2014), 80–109. https://doi.org/10.1016/j.destud.2013.10.002

[22] Kees Dorst and Nigel Cross. 2001. Creativity in the design process: co-evolution of problem–solution. *Design studies* 22, 5 (2001), 425–437. 02587.

[23] Claudia Eckert and Martin Stacey. 1998. Fortune favours only the prepared mind: Why sources of inspiration are essential for continuing creativity. *Creativity and Innovation Management* 7, 1 (1998), 1–12. 00056.

[24] C. Eckert and M. Stacey. 2003. Adaptation of Sources of Inspiration in Knitwear Design. *Creativity Research Journal* 15, 4 (2003), 355–384. 00000.

[25] Katherine Fu, Joel Chan, Jonathan Cagan, Kenneth Kotovsky, Christian Schunn, and Kristin Wood. 2013. The Meaning of Near and Far: The Impact of Structuring Design Databases and the Effect of Distance of Analogy on Design Output. *Journal of Mechanical Design* 135, 2 (2013), 021007. https://doi.org/10.1115/1.4023158

[26] William Gaver. 2011. Making spaces: how design workbooks work. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '11)*. Association for Computing Machinery, New York, NY, USA, 1551–1560. https://doi.org/10.1145/1978942.1979169

[27] D. Gentner and Arthur B. Markman. 1997. Structure mapping in analogy and similarity. *American Psychologist* 52, 1 (1997), 45–56.

[28] M. L. Gick and K. J. Holyoak. 1983. Schema induction and analogical transfer. *Cognitive Psychology* 15, 1 (1983), 1–38. https://doi.org/10.1016/0010-0285(83)90002-6 04045.

[29] M. M. Gielnik, M. Frese, J. M. Graf, and A. Kampschulte. 2011. Creativity in the opportunity identification process and the moderating effect of diversity of information. *Journal of Business Venturing* 27, 5 (2011), 559–576. https://doi.org/10.1016/j.jbusvent.2011.10.003 00000.

[30] Vinod Goel and Peter Pirolli. 1992. The structure of design problem spaces. *Cognitive Science* 16 (1992), 395–429.

[31] Milene Gonçalves, Carlos Cardoso, and Petra Badke-Schaub. 2013. Inspiration peak: exploring the semantic distance between design problem and textual inspirational stimuli. *International Journal of Design Creativity and Innovation* 1, 4 (2013), 215–232.

[32] Joy P. Guilford. 1967. *The nature of human intelligence*. McGraw-Hill, New York, NY.

[33] Andrew Hargadon and Robert I. Sutton. 1997. Technology Brokering and Innovation in a Product Development Firm. *Administrative Science Quarterly* 42, 4 (1997), 716–749. https://doi.org/10.2307/2393655 03534 Publisher: [Sage Publications, Inc., Johnson Graduate School of Management, Cornell University].

[34] M. E. Helms, S. Vattam, and A. K. Goel. 2008. Compound analogical design, or how to make a surfboard disappear. In *Proceedings of the 30th Annual Conference of the Cognitive Science Society*. 00013.

[35] Scarlett R. Herring, Chia-Chen Chang, Jesse Krantzler, and Brian P. Bailey. 2009. Getting inspired!: understanding how and why examples are used in creative design practice. In *Proceedings of the 27th international conference on Human factors in computing systems*. ACM, 87–96. https://doi.org/10.1145/1518701.1518717

[36] Thomas T. Hills, Peter M. Todd, David Lazer, A. David Redish, and Iain D. Couzin. 2015. Exploration versus exploitation in space, mind, and society. *Trends in Cognitive Sciences* 19, 1 (Jan. 2015), 46–54. https://doi.org/10.1016/j.tics.2014.10.004

[37] Keith J. Holyoak and Paul Thagard. 1996. *Mental leaps: Analogy in creative thought*. MIT Press, Cambridge, MA. 02398.

[38] Paul A. Howard-Jones, Sarah-Jayne . J. Blakemore, Elspeth A. Samuel, Ian R. Summers, and Guy Claxton. 2005. Semantic divergence and creative story generation: An fMRI investigation. *Cognitive Brain Research* 25, 1 (2005), 240–250.

[39] Alexander Ivanov, David Ledo, Tovi Grossman, George Fitzmaurice, and Fraser Anderson. 2022. MoodCubes: Immersive Spaces for Collecting, Discovering and Envisioning Inspiration Materials. In *Designing Interactive Systems Conference (DIS '22)*. Association for Computing Machinery, New York, NY, USA, 189–203. https://doi.org/10.1145/3532106.3533565

[40] David G. Jansson and Steven M. Smith. 1991. Design fixation. *Design Studies* 12, 1 (1991), 3–11.

[41] Eesh Kamrah, Fatemeh Ghoreishi, Zijian Ding, Joel Chan, and Mark Fuge. 2023. How Diverse Initial Samples Help and Hurt Bayesian Optimizers. *Journal of Mechanical Design* (July 2023), 1–32. https://doi.org/10.1115/1.4063006

[42] Youwen Kang, Zhida Sun, Sitong Wang, Zeyu Huang, Ziming Wu, and Xiaojuan Ma. 2021. MetaMap: Supporting Visual Metaphor Ideation through Multi-dimensional Example-based Exploration. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. ACM, Yokohama Japan, 1–15. https://doi.org/10.1145/3411764.3445325

[43] Craig A. Kaplan and Herbert A. Simon. 1990. In search of insight. *Cognitive Psychology* 22, 3 (1990), 374–419. https://doi.org/10.1016/0010-0285(90)90008-R

[44] Ianus Keller, Froukje Sleeswijk Visser, Remko van der Lugt, and Pieter Jan Stappers. 2009. Collecting with Cabinet: or how designers organise visual material, researched through an experiential prototype. *Design Studies* 30, 1 (Jan. 2009), 69–86. https://doi.org/10.1016/j.destud.2008.06.001

[45] G. Knoblich, S. Ohlsson, H. Haider, and D. Rhenius. 1999. Constraint relaxation and chunk decomposition in insight problem solving. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 25, 6 (1999), 1534–1555. https://doi.org/10.1037/0278-7393.25.6.1534 00691.

[46] Janin Koch, Andrés Lucero, Lena Hegemann, and Antti Oulasvirta. 2019. May AI?: Design Ideation with Cooperative Contextual Bandits. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (CHI '19)*. ACM, New York, NY, USA, 633:1–633:12. https://doi.org/10.1145/3290605.3300863 event-place: Glasgow, Scotland Uk.

[47] Janin Koch, Nicolas Taffin, Michel Beaudouin-Lafon, Markku Laine, Andrés Lucero, and Wendy E. Mackay. 2020. ImageSense: An Intelligent Collaborative Ideation Tool to Support Diverse Human-Computer Partnerships. *Proceedings of the ACM on Human-Computer Interaction* 4, CSCW1 (May 2020), 1–27. https://doi.org/10.1145/3392850 00010.

[48] Janin Koch, Nicolas Taffin, Andrés Lucero, and Wendy E. Mackay. 2020. SemanticCollage: Enriching Digital Mood Board Design with Semantic Labels. In *Proceedings of the 2020 ACM Designing Interactive Systems Conference*. ACM, Eindhoven Netherlands, 407–418. https://doi.org/10.1145/3357236.3395494

[49] Alex Kulesza and Ben Taskar. 2012. Determinantal point processes for machine learning. *Foundations and Trends® in Machine Learning* 5, 2-3 (2012), 123–286. https://doi.org/10.1561/2200000044

[50] Ranjitha Kumar, Jerry O. Talton, Salman Ahmad, and Scott R. Klemmer. 2011. Bricolage: example-based retargeting for web design. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '11)*. Association for Computing Machinery, New York, NY, USA, 2197–2206. https://doi.org/10.1145/1978942.1979262

[51] J. Richard Landis and Gary G. Koch. 1977. The Measurement of Observer Agreement for Categorical Data. *Biometrics* 33, 1 (March 1977), 159. https://doi.org/10.2307/2529310

[52] Jillian E. Lauer, Eukyung Yhang, and Stella F. Lourenco. 2019. The development of gender differences in spatial reasoning: A meta-analytic review. *Psychological Bulletin* 145 (2019), 537–565. https://doi.org/10.1037/bul0000191 Place: US Publisher: American Psychological Association.

[53] Brian Lee, Savil Srivastava, Ranjitha Kumar, Ronen Brafman, and Scott R. Klemmer. 2010. Designing with Interactive Example Galleries. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '10)*. ACM, New York, NY, USA, 2257–2266. https://doi.org/10.1145/1753326.1753667

[54] Julie S. Linsey, Ian Tseng, Katherine Fu, Jonathan Cagan, Kristin L. Wood, and Christian D. Schunn. 2010. A Study of Design Fixation, Its Mitigation and Perception in Engineering Design Faculty. *Journal of Mechanical Design* 132, 4 (2010), 041003.

[55] Andrés Lucero. 2012. Framing, aligning, paradoxing, abstracting, and directing: how design mood boards work. In *Proceedings of the Designing Interactive Systems Conference (DIS '12)*. Association for Computing Machinery, New York, NY, USA, 438–447. https://doi.org/10.1145/2317956.2318021

[56] Nic Lupfer, Andruid Kerne, Andrew M. Webb, and Rhema Linder. 2016. Patterns of Free-form Curation: Visual Thinking with Web Content. In *Proceedings of the 2016 ACM on Multimedia Conference (MM '16)*. ACM, New York, NY, USA, 12–21.

https://doi.org/10.1145/2964284.2964303
[57] Stephen MacNeil, Zijian Ding, Kexin Quan, Ziheng Huang, Kenneth Chen, and Steven P. Dow. 2021. ProbMap: Automatically constructing design galleries through feature extraction and semantic clustering. In *The Adjunct Publication of the 34th Annual ACM Symposium on User Interface Software and Technology*. ACM, Virtual Event USA, 134–136. https://doi.org/10.1145/3474349.3480203 00000.
[58] Stephen MacNeil, Ziheng Huang, Kenneth Chen, Zijian Ding, Alex Yu, Kendall Nakai, and Steven P. Dow. 2023. Freeform Templates: Combining Freeform Curation with Structured Templates. In *Creativity and Cognition*. 478–488. https://doi.org/10.1145/3591196.3593337 arXiv:2305.00937 [cs].
[59] N. R. F. Maier. 1931. Reasoning in humans. II. The solution of a problem and its appearance in consciousness. *Journal of Comparative Psychology* 12, 2 (1931), 181–194. https://doi.org/10.1037/h0071361
[60] R. L. Marsh and G. H. Bower. 1993. Eliciting cryptomnesia: unconscious plagiarism in a puzzle task. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 19, 3 (1993), 673–88.
[61] R. L. Marsh, T. B. Ward, and J. D. Landau. 1999. The inadvertent use of prior knowledge in a generative cognitive task. *Memory & Cognition* 27, 1 (1999), 94–105. https://doi.org/10.3758/bf03201216
[62] Winter Mason and Duncan J. Watts. 2012. Collaborative learning in networks. *Proceedings of the National Academy of Sciences* 109, 3 (Jan. 2012), 764–769. https://doi.org/10.1073/pnas.1110069108
[63] T. P. McNamara. 1992. Priming and Constraints It Places on Theories of Memory and Retrieval. *Psychological Review* 99, 4 (1992), 650–662. https://doi.org/10.1037/0033-295x.99.4.650
[64] Philip Mendels, Joep Frens, and Kees Overbeeke. 2011. Freed: a system for creating multiple views of a digital collection during the design process. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, Vancouver BC Canada, 1481–1490. https://doi.org/10.1145/1978942.1979160
[65] A. Newell and H. A. Simon. 1972. *Human problem solving*. Englewood Cliffs, NJ.
[66] Bernard A. Nijstad and Wolfgang Stroebe. 2006. How the group affects the mind: a cognitive model of idea generation in groups. *Personality and Social Psychology Review* 10, 3 (2006), 186–213. https://doi.org/10.1207/s15327957pspr1003_1 00643.
[67] T. Okada, S. Yokochi, K. Ishibashi, and K. Ueda. 2009. Analogical modification in the creation of contemporary art. *Cognitive Systems Research* 10, 3 (2009), 189–203. 00032.
[68] Srishti Palani, Zijian Ding, Stephen MacNeil, and Steven P. Dow. 2021. The "Active Search" Hypothesis: How Search Strategies Relate to Creative Learning. In *Proceedings of the 2021 Conference on Human Information Interaction and Retrieval*. ACM, Canberra ACT Australia, 325–329. https://doi.org/10.1145/3406522.3446046
[69] Srishti Palani, Zijian Ding, Austin Nguyen, Andrew Chuang, Stephen MacNeil, and Steven P. Dow. 2021. CoNotate: Suggesting Queries Based on Notes Promotes Knowledge Discovery. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems (CHI '21)*. Association for Computing Machinery, New York, NY, USA, 1–14. https://doi.org/10.1145/3411764.3445618
[70] Amy Pavel, Floraine Berthouzoz, and Bjorn Hartmann. [n.d.]. Browsing and Analyzing the Command-Level Structure of Large Collections of Image Manipulation Tutorials. ([n. d.]), 12.
[71] David N Perkins, David N. Perkins, and Margaret A Boden. 1994. Creativity: beyond the Darwinian paradigm. In *Dimensions of creativity*. MIT Press, Cambridge, MA, 119–142.
[72] P. Pirolli and S. Card. 1999. Information foraging. *Psychological Review* 106, 4 (1999), 643–675. https://doi.org/10.1037/0033-295x.106.4.643
[73] Jeroen G. Raaijmakers and Richard M. Shiffrin. 1981. Search of associative memory. *Psychological Review* 88, 2 (1981), 93–134.
[74] Samuel Rhys Cox, Yunlong Wang, Ashraf Abdul, Christian von der Weth, and Brian Y. Lim. 2021. Directed Diversity: Leveraging Language Embedding Distances for Collective Creativity in Crowd Ideation. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems (CHI '21)*. Association for Computing Machinery, New York, NY, USA, 1–35. https://doi.org/10.1145/3411764.3445782
[75] Mark A. Runco. 2010. Divergent thinking, creativity, and ideation. In *The Cambridge handbook of creativity*. Cambridge University Press, New York, NY, US, 413–446. https://doi.org/10.1017/CBO9780511763205.026
[76] Daniel M Russell, Malcolm Slaney, Yan Qu, and Mave Houston. 2006. Being literate with large document collections: Observational studies and cost structure tradeoffs. In *System Sciences, 2006. HICSS'06. Proceedings of the 39th Annual Hawaii International Conference on*, Vol. 3. IEEE, 55–55.
[77] Daniel M. Russell, Mark J. Stefik, Peter Pirolli, and Stuart K. Card. 1993. The Cost Structure of Sensemaking. In *Proceedings of the INTERACT '93 and CHI '93 Conference on Human Factors in Computing Systems (CHI '93)*. ACM, New York, NY, USA, 269–276. https://doi.org/10.1145/169059.169209
[78] Moushumi Sharmin and Brian P. Bailey. 2013. ReflectionSpace: an interactive visualization tool for supporting reflection-on-action in design. In *Proceedings of the 9th ACM Conference on Creativity & Cognition (C&C '13)*. Association for Computing Machinery, New York, NY, USA, 83–92. https://doi.org/10.1145/2466627.2466645
[79] Moushumi Sharmin, Brian P. Bailey, Cole Coats, and Kevin Hamilton. 2009. Understanding knowledge management practices for early design activity and its implications for reuse. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, Boston MA USA, 2367–2376. https://doi.org/10.1145/1518701.1519064 00047.
[80] Pao Siangliulue, Kenneth C Arnold, Krzysztof Z Gajos, and Steven P Dow. 2015. Toward collaborative ideation at scale: Leveraging ideas from others to generate more creative and diverse ideas. In *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing*. 937–945.
[81] Pao Siangliulue, Joel Chan, Steven P. Dow, and Krzysztof Z. Gajos. 2016. IdeaHound: Improving Large-scale Collaborative Ideation with Crowd-Powered Real-time Semantic Modeling. In *Proceedings of the 29th Annual Symposium on User Interface Software and Technology (UIST '16)*. ACM, New York, NY, USA, 609–624. https://doi.org/10.1145/2984511.2984578
[82] Pao Siangliulue, Joel Chan, Krzysztof Gajos, and Steven P. Dow. 2015. Providing timely examples improves the quantity and quality of generated ideas. In *Proceedings of the ACM Conference on Creativity and Cognition*. https://doi.org/10.1145/2757226.2757230
[83] Ut Na Sio, Kenneth Kotovsky, and Jonathan Cagan. 2015. Fixation or inspiration? A meta-analytic review of the role of examples on design processes. *Design Studies* 39 (July 2015), 70–99. https://doi.org/10.1016/j.destud.2015.04.004 00174.
[84] Steven M. Smith, T. B. Ward, and Jay S. Schumacher. 1993. Constraining effects of examples in a creative generation task. *Memory & Cognition* 21, 6 (1993), 837–45.
[85] Luis A. Vasconcelos, Maria A. Neroni, and Nathan Crilly. 2018. The effect of explicit instructions in idea generation studies. *AI EDAM* 32, 3 (Aug. 2018), 308–320. https://doi.org/10.1017/S0890060417000658 00009 Publisher: Cambridge University Press.
[86] Shaun Wallace, Brendan Le, Luis A. Leiva, Aman Haq, Ari Kintisch, Gabrielle Bufrem, Linda Chang, and Jeff Huang. 2020. Sketchy: Drawing Inspiration from the Crowd. *Proceedings of the ACM on Human-Computer Interaction* 4, CSCW2 (Oct. 2020), 1–27. https://doi.org/10.1145/3415243
[87] Hao-Chuan Wang, Dan Cosley, and Susan R. Fussell. 2010. Idea expander: supporting group brainstorming with conversationally triggered visual thinking stimuli. In *Proceedings of the 2010 ACM conference on Computer supported cooperative work - CSCW '10*. ACM Press, Savannah, Georgia, USA, 103. https://doi.org/10.1145/1718918.1718938
[88] Yunlong Wang, Priyadarshini Venkatesh, and Brian Y Lim. 2022. Interpretable Directed Diversity: Leveraging Model Explanations for Iterative Crowd Ideation. In *CHI Conference on Human Factors in Computing Systems*. ACM, New Orleans LA USA, 1–28. https://doi.org/10.1145/3491102.3517551
[89] T. B. Ward. 1994. Structured imagination: The role of category structure in exemplar generation. *Cognitive Psychology* 27, 1 (1994), 1–40. 00000.
[90] T. B. Ward. 1998. Analogical distance and purpose in creative thought: Mental leaps versus mental hops. In *Advances in Analogy Research: Integration of Theory and Data from the Cognitive, Computational, and Neural Sciences*, Keith J. Holyoak, Dedre Gentner, and B. Kokinov (Eds.). Sofia, Bulgaria, 221–230. 00142.
[91] Andrew M. Webb, Andruid Kerne, Rhema Linder, Nic Lupfer, Yin Qu, Kade Keith, Matthew Carrasco, and Yvonne Chen. 2016. A Free-Form Medium for Curating the Digital. In *Curating the Digital*. Springer, Cham, 73–87. https://doi.org/10.1007/978-3-319-28722-5_6
[92] M. J. Wilkenfeld and T. B. Ward. 2001. Similarity and Emergence in Conceptual Combination. *Journal of Memory and Language* 45, 1 (2001), 21–38. https://doi.org/10.1006/jmla.2000.2772
[93] Xiaotong (Tone) Xu, Rosaleen Xiong, Boyang Wang, David Min, and Steven P. Dow. 2021. IdeateRelate: An Examples Gallery That Helps Creators Explore Ideas in Relation to Their Own. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW2 (Oct. 2021), 1–18. https://doi.org/10.1145/3479496 00000.
[94] Huan Yuan, Kelong Lu, Mengsi Jing, Cuirong Yang, and Ning Hao. 2022. Examples in creative exhaustion: The role of example features and individual differences in creativity. *Personality and Individual Differences* 189 (April 2022), 111473. https://doi.org/10.1016/j.paid.2021.111473 00000.
[95] Darya L. Zabelina, Elina Zaonegina, William Revelle, and David M. Condon. 2021. Creative achievement and individual differences: Associations across and within the domains of creativity. *Psychology of Aesthetics, Creativity, and the Arts* (2021), No Pagination Specified–No Pagination Specified. https://doi.org/10.1037/aca0000439 Place: US Publisher: Educational Publishing Foundation.
[96] Liang Zeng, Robert W. Proctor, and Gavriel Salvendy. 2011. Fostering creativity in product and service development: validation in the domain of information technology. *Hum Factors* 53, 3 (2011), 245–70.
[97] Yinglong Zhang, Rob Capra, and Yuan Li. 2020. An In-situ Study of Information Needs in Design-related Creative Projects. In *Proceedings of the 2020 Conference on Human Information Interaction and Retrieval*. ACM, Vancouver BC Canada, 113–123. https://doi.org/10.1145/3343413.3377973 00003.

**APPENDIX**

# A PROCEDURE FOR GENERATING RUGGED WILDCAT WELLS SEARCH ENVIRONMENTS

We used four factors to control the synthetic objective functions:

(1) The *ruggedness amplitude*, in the range of [0,1], controlled the relative "height" of noise added to the search environment, compared to the height of the peaks. Increasing this parameter made the task more difficult by adding more places to incorrectly intuit as the location of the maximum reward. Setting this parameter to 1 would essentially make the search environment have infinite peaks.
(2) The *smoothness*, in the range of [0,1], controlled the degree of local correlation of scores in the grid. Intuitively, if smoothness was high (closer to 1), then a hill-climbing or gradient-based strategy could be viable, as a searcher that saw successive points in increasing score could (correctly) intuit that further points down that search path were likely to be of higher score; at lower values of smoothness, the search environment became very "bumpy", such that searchers would often be surprised by the score of nearby regions in the grid.
(3) The *number of peaks* controlled the number of "maxima" in the search environment; intuitively, this specified the number of regions in the space where a searcher might (correctly or otherwise) intuit as the location of the treasure. Mathematically, this parameter controlled the number of layers of multivariate normal with single peaks.
(4) The *distance between peaks*, in the range of [0,1], which prevented overlap of peaks when the function was generated with more than 1 peak.

All search environments in this study were generated with ruggedness amplitude set to 0.7 (fairly noisy), smoothness level to 0.2 (fairly rugged), and the number of peaks to 1 with a maximum of 100; distance between peaks was not relevant because we only used a single peak, using the following algorithm:

**Algorithm 1** Constructing the Wildcat Wells search environment with given ruggedness (noise) amplitude ($Rug_{amp}$), smoothness ($Smt$), number of peaks ($N$) and distance between peaks ($Rug_{freq}$).

1: **for** $Rug_{amp}, Smt, N, Rug_{freq}$ **do**
2: Get $X_{centers} = f(N, Rug_{freq})$
3: Sample $\sum_i^N surf \sim \mathbb{N}(X_i)$
4: Sample $Noise \sim$ OpenSimplex$(Smt)$
5: **end for**
6: **return** $surf + noise \times g(Rug_{amp})$.

# B ALGORITHM FOR SAMPLING DIVERSE AND NON-DIVERSE EXAMPLE POINTS

**Algorithm 2** Generating a ranked distribution of example sets with Determinantal Point Process (DPP) approach [49], $M$ is batch size (10000). $S^k$ is a combinatorial set defined on a finite set $X \in \mathbb{R}^2$, where each element $S^k_{Y_i} \in S^k$ is k elements long.

1: **for** $i \in range(M)$ **do**
2: Sample $S^k_{Y_i} \sim \mathbb{IID}(S^k)$ [identically sampling unordered sets without replacement]
3: Calculate $g(S^k_{Y_i}) = g_{y_i}$ and append this to $Scores_{S^k}$
4: **end for**
5: **return** DPP Score of sets of examples = $\frac{Score_{Sk} - mean(Score_{Sk})}{s.d.(Score_{Sk})}$.