Constructive Activities for People to Develop Their Creative Scientific Insights

ANONYMOUS AUTHOR(S)

Most people are different from institutional scientists in terms of lived experiences, objectives, and institutional knowledge. Such differences provide opportunities for *personal scientific creativity*: people can draw from—and contribute to—scientific knowledge beyond just donating data to experts. This paper studies creativity in scientific endeavors by non-scientists *when* supported with constructive activities. We provide empirical evaluation of two constructive techniques to support people in evaluating scientific explanations and designing experiments for a personal intuition. A between-subjects experiment tested whether asking readers to recreate an experiment leads them to focus more on underlying logic; participants asked to recreate explanations relied less on irrelevant surface details. A second between-subjects experiment tested whether support for procedural guidance assisted in experimental design; participants with access to procedural guidance created experimental designs that received higher scores from an experimental design expert. Our results suggest that constructive activities help people perform creative scientific work.

CCS Concepts: • **Human-centered computing → Empirical studies in HCI**.

Additional Key Words and Phrases: constructive activities, creativity, science, citizen science

1 INTRODUCTION

Most people are different from institutional scientists in terms of lived experiences, objectives, and institutional knowledge [15, 23]. Such differences provide opportunities and challenges for creative, complementary contributions to science. On the one hand, making scientific thinking accessible to people has multiple potential upsides: people can better understand scientific findings, apply those findings in their daily lives when appropriate, and contribute to expanding the scientific knowledge [15]. On the other hand, such contributions are one-off instances. While people have contributed important, diverse insights before [1, 31], the general lack of support for such *personal scientific creativity* is a missed opportunity for both science and for society.

This paper focuses on two common issues that show up when people perform scientific work: 1) focusing on surface-level details, and 2) difficulty getting started with scientific plans. The first concern is an instance of *fixation* where people focus extensively on some ideas while not considering a broader space of possibilities [29]. The second problem—lack of support for creating concrete artefacts—is a key challenge in creativity support. While prior research has extensively studied idea generation by novices [3, 28], research supporting people in implementing their ideas is relatively sparse.

This main contribution of this paper is a demonstration that constructive activities improve creative scientific work. *Constructive activities* "are those that require learners to produce some outputs, which may contained some new ideas, such as in self-explaining, drawing a concept map, or inducing hypotheses, and reflecting" [4]. We use ideas from constructive activities to improve people's performance on two scientific tasks—creating scientific explanations and designing an experiment for an intuition.

Manuscript submitted to ACM

1.1 Motivation: People demonstrate creative scientific insights but receive limited support

Laypeople insights have led to scientific findings [25, 31]. We focus on two aspects that enable this: a lack of expert blind spots, and unique lived experiences. People are often unencumbered by prior expert knowledge; sometimes, this can help them notice details experts might miss. E.g., lacking domain expertise about galaxies, citizen scientists on Galaxy Zoo fixated on green, blurry images while labeling galaxy images. Thus, a citizen science community *discovered* green pea galaxies [31]; experts had previously considered these galaxy images as apparatus error. Furthermore, unique lived experiences provide people different starting points to scientific enquiry than experts. Rather than starting with genetic data, users on the 23andme fora discuss their genetic testing results in terms of their and family members' traits and behavior. For instance, one person shared how they and their parents found chewing noises difficult to stand. Experts built on this insight to conduct further surveys and analysis which led to the discovery of misophonia markers [1]. In another popular domain—the microbiome—online citizen scientists' intuitions about their health and lifestyle provided novel intuitions to microbiome researchers [25]. This interpretive process of constructing personal knowledge is a form of personal or "mini-c" creativity that comes from the relationship of a person with their world [18].

As the green pea example highlights, focusing on surface features might sometimes provide insights that experts have overlooked; however, more often such *fixation* stands in the way of both creative ideation and understanding [29]. For tasks that require visual perception (e.g. labeling galaxy structures), fixation on surface features might help; in such cases, surface features–like the visual structure of a galaxy–contain valuable information. However, such fixation can cause misunderstandings when deeper knowledge is needed. For example, a 1993 study found that college students momentarily performed better on spatial reasoning when listening to Mozart [26]. The Mozart Effect paper only reported a temporary increase in spatial reasoning, yet numerous news articles claimed that Mozart makes people permanently "smarter". Compelling surface details like Mozart's name can overshadow the logic within an explanation.

1.2 Can constructive activities help people perform better scientific work?

Online citizens' intuitions about the microbiome provide another example of challenges faced by citizens. Early versions of "hypotheses" shared by users about the microbiome were closer to rambling accounts of personal health and lifestyle than falsifiable statements [24, 25]. Such accounts neither connected to existing science nor identified a possible relationship between independent and dependent variables. Just-in-time training for using scientific content and identifying hypotheses structure improved the quality of hypotheses created [25]. Such training is an instance of constructive support. Similarly, in the 23andme example, users generated the misophonia insight but did not validate it themselves; institutionally-trained scientists generated evidence for the insight. These instances highlight that converting an intuition/idea to an actual plan/artefact is difficult. The lack of support for transforming ideas into a concrete design is a critical problem in deepening and scaling creative work. Unsurprisingly, most citizen science projects support data collection efforts: experts decide what to track and people fill in the rows [32]. These projects assume that laypeople cannot quickly acquire sufficient domain knowledge to make useful contributions outside of data input. Might constructive activities help?

Constructive activities ask people to construct knowledge or inferences beyond the information directly given. In doing so, people produce and repair knowledge using the underlying structure of a situation rather than its surface details [5]. Being able to recognize which cues to focus on (and when) can likely improve scientific understanding and support complementary contributions. This paper examines whether constructive support can help novices in creative scientific activities. To reduce fixation on surface-level features in scientific explanations, participants recreated an experiment being explained. To transform intuitions to experimental designs, we provide a scaffolded approach with *procedural guidance* through examples, checklists, and templates. Our results demonstrate that constructive approaches reduce fixation on surface features and improve the quality of experimental designs.

2 RELATED WORK

We review research that describes why making creative scientific contributions is challenging, and provide some ideas to improve the current state.

2.1 Fixation on Surface Details Prevents Correct Evaluation but Recreation Might Help

Fixation hinders creative problem-solving by preventing a broad search of a solution space [29]. People might fixate by focusing on easily accessible surface-level features of an artefact. Such challenges show up in analogical thinking: participants do not successfully transfer analogies and focus on superficial feature(s) shared between the source and the target [17]. This focus on surface features shows up in multiple domains. While learning, novices commonly misunderstand explanations by overly relying on surface details—the literal objects, concepts, or entities explicitly described [7]—instead of evaluating underlying logic. In contrast, experts generally notice the deep structure of a situation that lies within their area of expertise [9]. While performing scientific tasks, participants fixated on surface details. Prior work has found adding a patina of neuroscience leads readers towards positively assessing explanations [34]. The study presented people with logically coherent and illogically circular science explanations. People generally perceived logical explanations as more satisfying than illogical ones. However, when irrelevant brain-related terminology was added to the explanations, novices in neuroscience rated illogical explanations more satisfying [34]. When evaluating a scientific explanation, fixation on surface details like neuroscience terminology may discourage people from examining the logic of the explanation.

One self-sourced approach to reduce fixation could be to understand the underlying logic of a process or an artefact. People may compare and contrast two scenarios [13] or self-explain a worked example [6]. More broadly, redoing someone else's work is a common learning strategy in creative disciplines, from painting to programming [10]. But people might lack the expertise to perform tasks for which they have limited or no training. One low-cost scaffold to induce such doing is perspective-taking. Perspective-taking has demonstrated improved diversity of ideas. In one version, participants asked to assume different roles generated more creative ideas[30]. For domain-specific tasks, performance of fixation fixes might depend on expertise too. For instance, one study found that experts were worse predictors of novice performance times, and resistant to debiasing techniques [16]. Since debiasing techniques worked better on those with lower expertise, maybe putting on an expert hat might help novices reduce fixation? Supporting perspective-taking with a recreation activity might be useful in reducing novices' fixation on surface features. This is the question for the first study in this work.

2.2 Construction Stalls at Idea Generation Due to Lack of Support; Guidance might help

Creativity support for domain-specific work beyond idea generation is missing. Prior work seeks to improve divergent thinking using examples and feedback: timely examples help [28], examples induce conformity [21], and combining reflection with feedback leads to extensive revisions [35]. While such techniques help with creating more/better ideas, support for creating domain-specific artefacts is far less common. For open-ended work like writing product reviews, Shepherd supports the creation process with expert feedback [11]; however, experts might not always be readily available to provide feedback. Furthermore, asking for inputs from experts, peers, or crowdworkers requires creating a first draft [14]. Creating the first design requires knowing the domain-specific rules and applying them correctly. This challenge extends to scientific work. Even when people come up with creative insights, evaluating them with scientific experiments is difficult. Scaffolding techniques–defined as "guiding individuals through smaller subtasks in sequence that, in turn, have them complete a larger complex task" [14]–have supported some scientific tasks before. Tummy trials supports people in setting up N=1 experiments by providing expert-curated choices to choose from. Foldit players demonstrates immense creativity in generating low-energy protein structures - a highly creative scientific activity [8]. Foldit achieves this by transforming a computational problem into a user-friendly game. Domain-specific protein folding knowledge is encoded in the procedural rules of the game where lower energy states (conceptually) correspond to better game scores. Unfortunately, such techniques are challenging for scientific tasks that are not as visually perceptive and highly structured as folding protein structures.

Experiment design exemplifies this challenge of providing general creativity support challenges for domain-specific work. Converting an intuition into an experimental design is difficult. First, it requires knowing the structure of an experiment. E.g. A between-subjects experiment design has a defined structure: a hypothesis, ind/dep vars, conditions, instructions. Second, making contextually-appropriate choices for these subparts require prior knowledge. E.g., knowing are the measures appropriate for the hypotheses? Third, experimental design is an iterative process; people learn about the constraints and expectations as they design the experiment. People need two kinds of support to perform complex new tasks: conceptual support (what to do), and procedural support (how to do it). For instance, when creating a new experimental design, this includes informational resources (what does an experiment contain?, how to create different parts of an experiment) and means to document their design. This necessitates explicit support for both the conceptual structure and the procedural steps to follow. Creative scientific work is a lot more open-ended; providing constructive scaffolds for different knowledge needs should help.

3 EXPERIMENT 1: DO CONSTRUCTIVE ACTIVITIES IMPROVE EXPLANATION COMPREHENSION?

This experiment compares a constructive activity with a recall activity for understanding science explanations. We hypothesized that compared to the recall activity, the creative task of recreating an explanation would reduce fixation on surface features. Additionally, we hypothesized that after recreating an explanation, participants will avoid fixating on neuroscience surface features in subsequent explanations.

3.1 Participants

Undergraduates were recruited from social science courses at a California research university (n = 72, 54 female). Participants received course credit for participation and were informed that the results of their experiment would have no impact on their class performance.

3.2 Design

Participants completed an online study with two tasks: *Comprehension* and *Ratings*. The first task—*Comprehension*—tests for the reliance on surface features.The second task–*Ratings*—tests the transfer of reduced reliance on surface features from the *Comprehension* task. There are two conditions for each task, resulting in a 2 x 2 design: (for Comprehension task) *Recreate* vs *Recall* × (for Ratings task) *Without Neuroscience* vs. *With Neuroscience*. We hypothesized that *Recreate* participants would avoid fixating on the surface details in the recreated explanation. We also hypothesized that *Recreate* participants would avoid fixating on surface details in subsequent explanations, even when not explicitly instructed to recreate them.

3.3 Materials and Procedure

In the *Comprehension* task, participants were shown a science explanation from a prior neuroscience study [34]. The *Recreate* group was asked to imagine themselves as scientists reconstructing the described experiment and answer questions about their results, while the *Recall* group was asked to recall answers from the given text (Table 1a). In the *Ratings* task, participants rated the quality of explanations copied from [34]. Each explanation either proposed a logical mechanism or provided a circular restatement of a psychology finding. A circular restatement (Table 1b) provides the same information as the description, not providing any additional explanatory power. Each subject rated 4 logical and 4 circular explanations in a random order. In the *With Neuroscience* condition, irrelevant neuroscience information was added to every explanation.

3.4 Measures

Independent variables are the *Comprehension* task questions (*Recall* vs. *Recreate*), *Ratings* explanation content (*With Neuro* vs. *Without Neuro*), and *Ratings* explanation quality (logical vs. circular). Dependent variables are the *Comprehension* responses and *Ratings* numeric ratings of explanation quality ranging from +3 (good) to -3 (bad). Two raters with scientific training and multiple peer-reviewed scientific publications independently rated 10 of the 72 entries, then discussed them to form a shared view of assessment. Next, each independently rated all 72 *Comprehension* responses on three binary scales: *Surface*, *True*, and *Alternative*.A response was marked as Surface if it referenced the irrelevant neuroscience information from the original explanation. A response was marked as True if it referenced the mechanism provided by the original explanation. Finally, a response was marked as Alternative if it proposed an alternative mechanism not directly present in the original explanation. The final score is the mean of the independent ratings for a feature. A high degree of reliability was found for neuro information. The average measure ICC was .805 with a 95% confidence interval from .704 to .874 ($F(70,70)= 10.8$, $p<.00001$). A medium degree of reliability was found for mechanism ratings. The average measure ICC was .472 with a 95% confidence interval from .27 to .635 ($F(70,70)= 2.09$, $p<.00001$). A high degree of reliability was found for inference ratings. The average measure ICC was .836 with a 95% confidence interval from .749 to .894 ($F(70,70)= 8.06$, $p<.00001$).

Table 1. Tasks in Experiment 1: (a) *Comprehension* task: Explanation and questions for each condition. Participants also read a description of the experiment (not shown). (b) *Ratings* task: Example of a circular explanation with neuroscience.

(a) ***Comprehension*** **task**	(b) ***Ratings*** **task**
Explanation: Information about stereotypical animals is stored in a certain way by CA3 brain cells, which have been shown to mediate memory. This makes the information more readily accessed and manipulated than information about rare animals.	**Description** (excerpt): The researchers discovered that words spoken soon after a presented target word were words that sounded like the target, while words spoken later were words that had a similar meaning to the target.
Recall: Based on the explanation above, why was one type of animal easier to reason about than another? *Recreate*: Suppose you are a scientist recreating this experiment and find similar results. Why might your subjects be better at reasoning about stereotypical animals than rare animals?	Rate the quality of the following explanation: Patterns of brain activation in these subjects lead researchers to conclude that this happens because Broca's area, a part of the brain's language system, associates two different types of words with the target word at two different times.

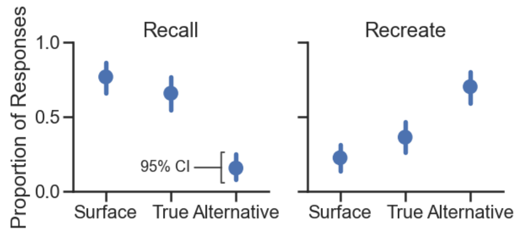


Fig. 1. In the *Comprehension* task, *Recall* participants utilize irrelevant neuroscience information, while *Recreate* participants propose alternative mechanisms without relying on neuroscience information.

3.5 Results

3.5.1 Recreate Participants Rely Less on Surface Features and Generate Alternative Mechanisms. In the *Comprehension* task, *Recall* participants relied on the explanation's text. When asked why an experimental finding occurred, they often included the explanation's provided mechanism but also its irrelevant neuroscience information (Figure 1). Compared to *Recall*, *Recreate* participants were less likely to include the explanation's mechanism and neuroscience information (True: $t(135.7) = 3.64$, $p < 0.01$; Surface: $t(134.2) = 7.60$, $p < 0.01$). We reflect on this in the discussion section. *Recreate* participants generated alternative mechanisms more often than *Recall* (Alternative: $t(142.0) = -7.89$, $p < 0.01$).

3.5.2 Neuroscience Detail Increases Ratings of Circular Explanations. Contrary to our hypothesis, participants in both conditions rated circular explanations with neuroscience higher quality than circular explanations without neuroscience (*Recall*: $t(125.6) = -2.10$, $p < 0.05$; *Recreate*: $t(157.2) = -3.233$, $p < 0.01$) (Figure 2a). In addition, there was no significant difference between *Recall* and *Recreate* for ratings of circular explanations with neuroscience ($t(121.8) = -1.8$, $p > 0.05$). These patterns are consistent with prior work that did not include a task before ratings, suggesting that neither *Recall* nor *Recreate* mitigated the positive bias caused by neuroscience surface details. When explanations did not include neuroscience, participants rated logical explanations higher quality than circular explanations (*Recall*: $t(119.1) = 5.69$, $p < 0.01$; *Recreate*: $t(180.4) = 3.22$, $p < 0.01$) (Figure 2b). This is also consistent with prior work, suggesting that participants perceived a difference between logical and circular explanations when neuroscience was not included.

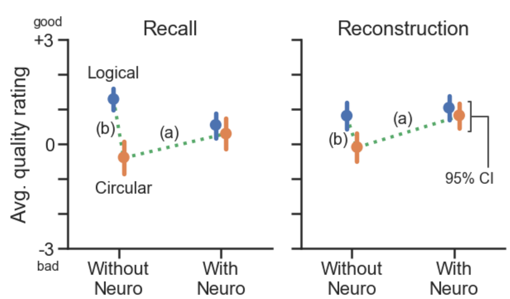


Fig. 2. Contrary to our hypothesis, participants in both conditions rated circular explanations higher quality when explanations contained neuroscience information (a). Participants rated circular explanations without neuroscience lower quality than logical explanations (b).

3.6 Discussion

3.6.1 Recreation Reduced Fixation on Irrelevant Surface Details. 59.4% of participants in the Recall condition made reference to the neuroscience information in their response while 20% of participants in the Recreate condition made reference to the neuroscience information. Qualitatively, a number of responses in the Recall condition were nearly word-for-word copies of the explanation text. Despite the inclusion of an irrelevant neuroscience surface detail in the explanation, *Recreate* participants seldom used this detail when recreating the explanation. In fact, *Recreate* participants proposed alternative mechanisms instead of referencing either neuroscience detail or original mechanism in their explanations. One interpretation is that these participants ignored the text entirely or did not understand the mechanism. If this was the case, the proposed mechanisms should be irrelevant or inconsistent. Most *Recreate* participants, however, proposed relevant mechanisms that accounted for the specific experimental results discussed in the text. Some proposed mechanisms referenced the original explanation and elaborated on it, demonstrating both knowledge of and ability to extend the text's structure. While *Recall* responses were often word-for-word copies of the text, *Recreate* responses proposed a variety of mechanisms.

3.6.2 Recreation participants proposed creative explanations. Participants possibly drew from some prior scientific knowledge; they used terms suggesting mechanisms based on prior knowledge or "confidence"; one explicitly used the scientific term "schema" (highlighted by the lead author).

"People are more likely to have **previous knowledge** on stereotypical birds which makes it easier to understand the new information."

"It may be possible that not only is this information easier to access, but participants are more **confident** because of their familiarity with stereotypical birds."

"human beings already have **certain schemas** that help them make sense of the world. stereotypical animals are more accessible schemas and are more commonly referenced in everyday life, as opposed to rare animals."

Not all *Recreate* participants' explanations were useful. Some participants proposed more outlandish theories.

"a lot of the people like to follow the norm so they follow within certain social guidelines that may be stereotypical"

"subjects are better at reasoning about stereotypical animals because they have been exposed to some type of information by society. they have somewhat of an understanding of how these stereotypical animals are seen through experiences of others."

3.6.3 Recreation participants displayed more words suggesting role-taking. Additionally, some Recreate participants used hypothetical language suggesting they slipped into the role of a scientist; participants used words like "subjects"/"my subjects"/ "i believe" and others demonstrating lack of surity, such as "probably" /"they might"/ "likely to have". These patterns might have come from people making creative guesses while recreating the experiment; future work can investigate such questions.

"**my subjects** may be better at reasoning about stereotypical animals than rare animals because stereotypical animals are more easily accessible in the way they are stored by ca3 brain cells"

"in **my version**, i would simplify the study"

This experiment demonstrated that people people reduce fixation on surface features when prompted with recreating the explanation as a scientists. People generated multiple alternative explanations; such explanations are crude hypotheses that people could test with some support for experiment design. The second experiment describes this activity—designing experiments for personal intuitions— that is personally motivating, but also requires more knowledge support.

4 EXPERIMENT 2: EXPERIMENT DESIGN WITH PROCEDURAL GUIDANCE

Experiment design presents many challenges. For instance, a person needs to know the structure of an experiment to create one. A between-subjects experiment design has a defined structure: a hypothesis, ind/dep vars, conditions, instructions. Many online videos provide definitions and conceptual knowledge; however, they do not provide how-to resources for creating such structure [19]. Templates for between-subjects experiments might help people convert an intuition into the structure of an experiment (e.g. starting by converting an intuition to a hypothesis). However, people would still need help filling in the different parts of this structure; we call support for filling in the different components–using examples, checklists, templated options–as *procedural guidance*. We hypothesized that participants who use procedural guidance create better experiment designs than those who watch videos on the topic. A between-subjects experiment tested this hypothesis.

4.1 Method and Design

The study asked participants to compose an experimental design for a personal intuition of their choosing. Participants were randomly assigned to one of two conditions: Tutorial or Procedural Guidance (PG) (Figure 3). Each condition provided informational resources and a means to document their design (Tutorial with a text document, or procedural guidance with inline text fields). Moreover, participants were provided instructions that the resources described the attributes that their designs should possess. Scripted study instructions ensured the same manipulation between the two conditions.

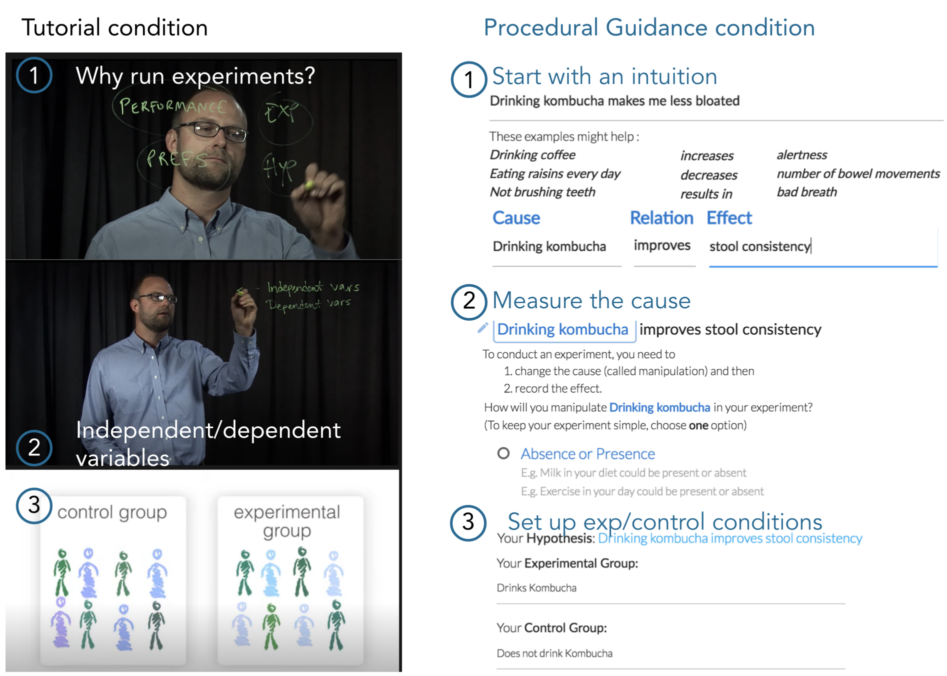


Fig. 3. In Study 2, two conditions offered equivalent content through different means. The Tutorial condition (left) provided short, topical videos; for each video, the number and the text present the order and topic of the video. The Procedural Guidance condition (right) provided examples, checklists, and templated options for different steps of designing an experiment.

The Tutorial condition provided a playlist of six videos about experiment design (mean length: 3min 30sec). All videos, except one, showed an expert in experiment design define and provide details about different experiment components; three are shown in Figure 3. These videos were curated from a MOOC about designing and running experiments. One video (about experimental and control conditions) was sourced from a Clinical and Translational Research Institute (CTRI) at an American public University. The videos were lightly edited to focus on material relevant to designing an experiment. The Procedural Guidance (PG) condition provided participants access to similar information about experiment design. In this condition, participants followed a guided interface that displayed examples, checklists, and templated options.

Both conditions had access to similar content for creating a structurally-sound experiment; they differed in the nature of support in two key ways: 1) *just-in-time*: In the PG condition, participants received appropriate examples and other support only upon reaching that step. In the Videos condition, participants were not restricted from exploring any video at any time without having to create a design. 2) *in-situ*: In the PG condition, people received 'how-to' help (examples, templates, and checklists) in the same interface that they created the experiment design in. In the Tutorial condition, participants saw the videos on a browser tab and typed in a google doc in a separate tab.

Participants were told that there was no lower- or upper-time limit on how long they took on the task. Each session comprised the following steps: consent, design task, survey, and interview. Participants could also use web resources–such as Wikipedia–and many did. The interview asked participants about confidence in their experiment design abilities and their experience using the system. The interview was tailored to participants' behavior and survey responses: for example, if a participant did not watch some videos, the interviewer asked why. An independent rater (a professor who teaches experiment design) blind to condition rated each participant's experiment using the rubric.

4.2 Participants

Recruitment: 72 participants were recruited from a Western US Research University (Table 2). 11 had no prior experience with experiment design; 61 had taken a course or equivalent. Expertise was counterbalanced across conditions.

4.3 Measures

The study scored experiments via a 13-question rubric (Table 3a), and recorded time taken. A blind-to-condition expert (a regular instructor of large, undergraduate courses on experiment design) provided the scores for the experiment design. The rubric was developed iteratively by the lead author & an instructor (an expert in research methods instruction)

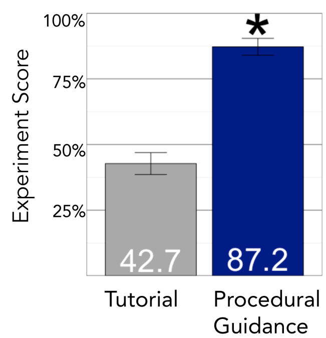
Table 2. Demography info for 72 participants (all undergraduate students). Some participants did not complete portions of the survey.

Nationality	USA = 37	China = 11
	No Answer = 6	Others = 18
Gender	Female = 47	Male = 24
Native English	Yes = 38	No = 34
Age	18-20 = 40	26-30 = 1
	21-25 = 31	
Ethnicity	Asian/Pacific = 36	Hispanic/Latino = 14
	White = 11	Others = 11
Major	Biology = 12	Psychology = 20
	Cognitive Sci = 12	Others = 20
Used online learning	Never = 28	Occasional = 16
	1 class = 11	2-5 classes = 12

Table 3. Details for Experiment 2:
a) Measure: Rubric for design-quality criteria for Structure (13 points)
b) Result: Access to Procedural Guidance improved the quality of experiment design. Mann–Whitney U = 108; n1 = n2 = 36, $p < 0.005$.

a
Hypothesis: 3 points
Is the cause/relation/effect specific? (1pt each)
Measurement: 2 points
Are the cause and effect manipulated/measured correctly? (1pt each)
Conditions: 3 points
Are the control and experimental conditions appropriate? 2pts
Do the conditions differ in manipulating the cause? 1pt
Steps: 2 points
Are experimental steps clear for control/experimental conditions?
Criteria: 2 points
Are the exclusion criteria correct and complete?
Are the inclusion criteria correct?
Can the overall experiment be run as-is? 1 point

b



during an early pilot in a class. The rubric checks whether people create correct specific elements of an experiment. Qualitative measures included how participants used the tool, where they faced challenges, and a post-experiment survey. A non-parametric Mann-Whitney test assessed the effect of condition on design quality.

4.4 Results

Participants in the PG condition created higher-quality experiments (M = 11.3) than Tutorial participants (M = 5.6); Mann–Whitney U = 108, n1 = n2 = 36, $p < 0.005$ (Table 3b). Of the 36 designs rated in the top half, 29 were from PG condition. PG participants performed better on five out of six sections (all except hypothesis). There was no significant difference in the amount of time participants spent creating an experiment in the Tutorial (M = 30.8 mins) vs PG (M = 29.0 mins) conditions; Mann–Whitney U = 734, n1 = n2 = 36, p = 0.33 two-sided.

4.5 Discussion

As PG aims to improve creative knowledge work, like experimental design, the primary dependent variable was the quality of the experiment design. Online video resources–as provided in the Tutorial condition–represent a common status quo: contemporary and bite-sized yet still static resources. This comparison enabled us to observe how procedural support changed design outcomes compared to a common way people consume (educational) information online.

4.5.1 Why did participants with procedural guidance design structurally-sound experiments? Procedural Guidance participants performed better on all aspects of experiment design and produced more high-scoring experimental designs (Figure 4). Tutorial condition's lower score and our observations suggest contextually-integrated approaches like procedural support increase useful adoption of information due to three reasons. First, receiving a structure provides a headstart. Tutorial participants wrote down specific words like "independent variable" and "dependent variable" in their sheet to to fill in later. Most Tutorial participants used topics and keywords from videos to structure their experiment. Second, in-situ support helps. PG participants mentioned that the interface provided sufficient examples. Participants in the Tutorial condition felt that the videos provided a refresher of some concepts they vaguely knew about but that the videos felt slow. Tutorial participants followed one of two strategies: 1) watch all the videos at once and then

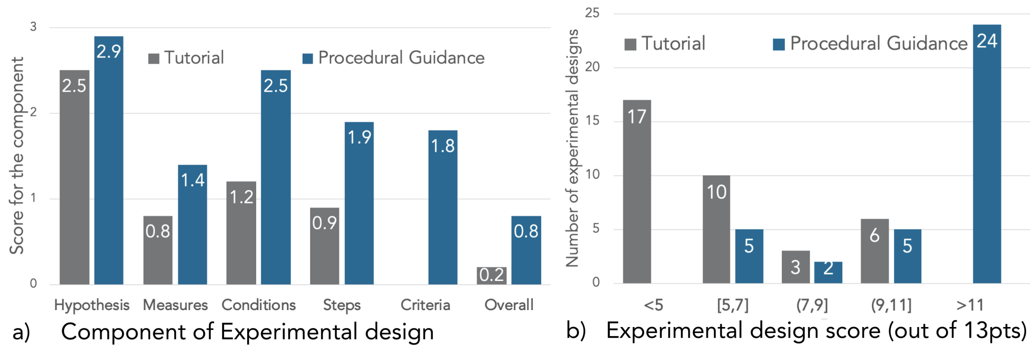


Fig. 4. a) PG experiments' components scored higher than Tutorial experiments. b) Most Tutorial designs (27/36) scored less than 7 points, while most PG designs (31/36) scores more than 7 points.

begin writing the experiment; or 2) begin designing the experiment and use the videos to fill in the gap when stuck. Like cramming, all-at-once watching floods the mind, perhaps making it difficult to use seen ideas [20]. By contrast, the search-when-needed approach interrupts flow, replacing the attention on design with a task of locating needed information. Third, people successfully translate examples to their setting. We provide one comparison: Participants in both conditions verbally expressed a lack of confidence in their chosen cause/effect measures; PG participants demonstrated high scores for the section but Tutorial folks did not. PG participants had access to templated options for measurements that many reused; Tutorial participants did not have this option. Furthermore, during interviews, Tutorial participants wanted to see recommendations about how past experiments have measured the variables they are interested in; many explicitly mentioned the need for more examples.

4.5.2 People made creative choices and drew from personal experience. People made surprising, creative choices in the content of the experiments, sometimes spending substantial time. Many participants searched online to find technical details and measures. Some spent over 15 minutes searching online for measures: one found a formal sleep-quality scale from Stanford researchers. People found online resources of varying utility. E.g., using JOVE, a database for peer-reviewed scientific video protocols and tracking REM sleep quality (mattressadvisor.com/rem-sleep) might be useful for a sleep-related experiment but other choices–like finding measures for keratinocyte production–were less relevant. Participants in both conditions mentioned that they enjoyed reflecting on their lifestyle/health ideas and thinking through how to transform an intuition into an experiment. Participants wished that the tool was integrated with their class, describing it as "hands on" and "DIY."

Experiment designs showcased topics of personal interest. Majority of experiments were about sleep combined with topics from personal health and performance. While the focus on sleep could represent some fixation with the given example note about sleep, people still had a variety of questions about sleep. E.g., "*I am more awake and energetic in the morning if I am woken up abruptly by an alarm or a person* (S12)" or "*I feel more awake when I take my iron supplement* (S33)" Participants mentioned that their intuitions were based on personal curiosity; e.g. "*Does physical activity help reduce anxiety and stress levels?* (S49)" and "*I am more sluggish throughout the day if I hit the snooze button* (S57)". Most designs demonstrated a topic that was discussed on other online fora, showing that others have similar intuitions about health; e.g., "*Exercising right before I fall asleep makes it much harder for me to fall asleep* (S54)".

5 GENERAL DISCUSSION

In this section, we discuss and synthesize the findings from our work

5.1 Supporting Personal Scientific Creativity

Creativity is not just about significant novel invention with large social impact. We believe personal creativity is of equal interest to the Creativity & Cognition community and of benefit to human well-being. For instance, lead users create novel designs from existing products for their needs; such designs benefit many [33]. Therefore, supporting people in converting their creative ideas to designs is likely a worthwhile goal. Both the studies demonstrated creative insights from participants: alternative mechanisms in the first experiment and creative personal intuitions for experimentation in the second. Both these represent a form of personal, or mini-c creativity [18].

Better understanding scientific thinking can also yield promising hypotheses for human cognition. E.g., prior work has studied scientists with the in-vivo/in-vitro model of first observing and analyzing scientists at laboratory meetings, followed by in-vitro lab experiments on the cognitive processes identified [12]. Further, improving people's scientific work could help draw insights for other disciplines–like design–that also follow an iterative, structured process. We believe our work brings some attention to the thinking and doing of science by novices; we hope future research will build on such ideas. One limitation of our work is the participant set: students at a public university. Our participants might be more informed on experiments than someone without a college degree. At the same time, students might also be less motivated than some communities with more important personal needs. Citizen publics discuss and use institutional knowledge [22]; however, general uptake of such conversation might be uneven [23]. People fall along a broad spectrum in terms of their involvement with science: lead users, self-tracking enthusiasts (like the quantified self community), self-experimenters (like kombucha brewers), informed citizens, those with advanced degrees in science, and more. While many citizens are intrigued by scientific results and papers, some might find the structured nature of scientific to be flat and formal. Despite such limitations, supporting more people in performing creative scientific work would be an improvement on the status quo.

5.2 Creating Techniques for Deeper Creative Production

Design, creativity, and crowdsourcing researchers have evaluated many techniques to improve the quantity and quality of ideas generated by novices. Multiple interventions have successfully improved the diversity of ideas, like providing timely examples [28], task-specific feedback [11], and providing explanations with ideas [2]. Perspective-taking can also encourage people to explore more alternatives [30]. Generating more ideas, however, is one half of the creative process for domain-specific work; eventually, the idea needs to be converted to an artefact that can be evaluated. Our experience provides some insights on designing systems/techniques that support converting ideas to artefacts. One way to support a complex task is to create an appropriate activity structure that naturally translates to pedagogical tools like procedural guidance. We found that in-situ, timely examples helped. Constructing own knowledge by can help prime people to learn better; this is especially true for novices who, unlike experts, might not have prior knowledge to activate [27]. Prior research has noted the same; good support systems are "by default, task-specific and specific step-specific" [11]. Extending these findings, procedural guidance draws from the same hypothetical experiment to share related examples across different steps. Such continuous guidance can make examples more direct and useful for people as they transform their ideas to designs.

REFERENCES

[1] 23andMe. 2016. Something to Chew On. https://blog.23andme.com/23andmeresearch/something-to-chew-on/. (2016). https://blog.23andme.com/23andmeresearch/something-to-chew-on/

[2] Faez Ahmed, Nischal Reddy Chandra, Mark Fuge, and Steven Dow. 2019. Structuring Online Dyads: Explanations Improve Creativity, Chats Lead to Convergence. In *Proceedings of the 2019 on Creativity and Cognition*. 306–318.

[3] 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*. 93–105.

[4] Michelene TH Chi. 2009a. Active-constructive-interactive: A conceptual framework for differentiating learning activities. *Topics in cognitive science* 1, 1 (2009), 73–105.

[5] Michelene TH Chi. 2009b. Active-Constructive-Interactive: A Conceptual Framework for Differentiating Learning Activities. *Topics in cognitive science* 1, 1 (2009), 73–105.

[6] Michelene TH Chi, Miriam Bassok, Matthew W. Lewis, Peter Reimann, and Robert Glaser. 1989. Self-Explanations: How Students Study and Use Examples in Learning to Solve Problems. *Cognitive science* 13, 2 (1989), 145–182.

[7] Michelene TH Chi, Robert Glaser, and Ernest Rees. 1981. *Expertise in Problem Solving.* Technical Report. PITTSBURGH UNIV PA LEARNING RESEARCH AND DEVELOPMENT CENTER.

[8] Seth Cooper, Firas Khatib, Adrien Treuille, Janos Barbero, Jeehyung Lee, Michael Beenen, Andrew Leaver-Fay, David Baker, Zoran Popović, and others. 2010. Predicting protein structures with a multiplayer online game. *Nature* 466, 7307 (2010), 756–760.

[9] National Research Council and others. 1999. *How people learn: Bridging research and practice.* National Academies Press.

[10] Vanessa P Dennen and Kerry J Burner. 2008. The cognitive apprenticeship model in educational practice. *Handbook of research on educational communications and technology* 3 (2008), 425–439.

[11] Steven Dow, Anand Kulkarni, Scott Klemmer, and Björn Hartmann. 2012. Shepherding the crowd yields better work. In *Proceedings of the ACM 2012 conference on computer supported cooperative work*. 1013–1022.

[12] Kevin Dunbar. 2001. What scientific thinking reveals about the nature of cognition. *Designing for science: Implications from everyday, classroom, and professional settings* (2001), 115–140.

[13] Mary L. Gick and Keith J. Holyoak. 1983. Schema Induction and Analogical Transfer. *Cognitive psychology* 15, 1 (1983), 1–38.

[14] Michael D Greenberg, Matthew W Easterday, and Elizabeth M Gerber. 2015. Critiki: A scaffolded approach to gathering design feedback from paid crowdworkers. In *Proceedings of the 2015 ACM SIGCHI Conference on Creativity and Cognition*. 235–244.

[15] Susanne Hecker, Lisa Garbe, and Aletta Bonn. 2018. The European citizen science landscape–a snapshot. JSTOR.

[16] Pamela J Hinds. 1999. The curse of expertise: The effects of expertise and debiasing methods on prediction of novice performance. *Journal of experimental psychology: applied* 5, 2 (1999), 205.

[17] Keith James Holyoak and Robert G Morrison. 2005. *The Cambridge handbook of thinking and reasoning*. Vol. 137. Cambridge University Press Cambridge.

[18] James C. Kaufman. 2009. Beyond Big and Little : The Four C Model of Creativity.

[19] Juho Kim, Phu Tran Nguyen, Sarah Weir, Philip J Guo, Robert C Miller, and Krzysztof Z Gajos. 2014. Crowdsourcing step-by-step information extraction to enhance existing how-to videos. In *Proceedings of the SIGCHI conference on human factors in computing systems*. 4017–4026.

[20] Nate Kornell. 2009. Optimising learning using flashcards: Spacing is more effective than cramming. *Applied Cognitive Psychology: The Official Journal of the Society for Applied Research in Memory and Cognition* 23, 9 (2009), 1297–1317.

[21] Chinmay Kulkarni, Steven P Dow, and Scott R Klemmer. 2014. Early and repeated exposure to examples improves creative work. In *Design thinking research*. Springer, 49–62.

[22] Stacey Kuznetsov, Aniket Kittur, and Eric Paulos. 2015. Biological citizen publics: Personal genetics as a site of public engagement with science. In *Proceedings of the 2015 ACM SIGCHI Conference on Creativity and Cognition*. 303–312.

[23] Gwen Ottinger, D Tyfield, R Lave, S Randalls, and R Thorpe. 2017. Scientific authority and models of change in two traditions of citizen science. *The routledge handbook of the political economy of science* 351 (2017), 9781315685397–31.

[24] Vineet Pandey, Amnon Amir, Justine Debelius, Embriette R Hyde, Tomasz Kosciolek, Rob Knight, and Scott Klemmer. 2017. Gut instinct: Creating scientific theories with online learners. In *Proceedings of the 2017 CHI conference on human factors in computing systems*. 6825–6836.

[25] Vineet Pandey, Justine Debelius, Embriette R Hyde, Tomasz Kosciolek, Rob Knight, and Scott Klemmer. 2018. Docent: transforming personal intuitions to scientific hypotheses through content learning and process training. In *Proceedings of the Fifth Annual ACM Conference on Learning at Scale*. 1–10.

[26] Frances H. Rauscher, Gordon L. Shaw, and Catherine N. Ky. 1993. Music and Spatial Task Performance. *Nature* 365, 6447 (1993), 611.

[27] Daniel L Schwartz and John D Bransford. 1998. A time for telling. *Cognition and instruction* 16, 4 (1998), 475–5223.

[28] Pao Siangliulue, Joel Chan, Krzysztof Z Gajos, and Steven P Dow. 2015. Providing timely examples improves the quantity and quality of generated ideas. In *Proceedings of the 2015 ACM SIGCHI Conference on Creativity and Cognition*. 83–92.

[29] Steven M. Smith and Steven E. Blankenship. 1991. Incubation and the Persistence of Fixation in Problem Solving. *The American journal of psychology* (1991), 61–87.

[30] Jaime Teevan and Lisa Yu. 2017. Bringing the wisdom of the crowd to an individual by having the individual assume different roles. In *Proceedings of the 2017 ACM SIGCHI Conference on Creativity and Cognition*. 131–135.

[31] Ramine Tinati, Max Van Kleek, Elena Simperl, Markus Luczak-Rösch, Robert Simpson, and Nigel Shadbolt. 2015. Designing for citizen data analysis: A cross-sectional case study of a multi-domain citizen science platform. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*. 4069–4078.

[32] Rajan Vaish, Snehalkumar (Neil) S Gaikwad, Geza Kovacs, Andreas Veit, Ranjay Krishna, Imanol Arrieta Ibarra, Camelia Simoiu, Michael Wilber, Serge Belongie, Sharad Goel, and others. 2017. Crowd research: Open and scalable university laboratories. In *Proceedings of the 30th Annual ACM Symposium on User Interface Software and Technology*. 829–843.

[33] Eric Von Hippel. 2006. *Democratizing innovation*. the MIT Press.

[34] Deena Skolnick Weisberg, Frank C. Keil, Joshua Goodstein, Elizabeth Rawson, and Jeremy R. Gray. 2008. The Seductive Allure of Neuroscience Explanations. *Journal of cognitive neuroscience* 20, 3 (2008), 470–477.

[35] Yu-Chun Grace Yen, Steven P Dow, Elizabeth Gerber, and Brian P Bailey. 2017. Listen to others, listen to yourself: Combining feedback review and reflection to improve iterative design. In *Proceedings of the 2017 ACM SIGCHI Conference on Creativity and Cognition*. 158–170.