Submitted to *Manufacturing & Service Operations Management*

On Repeat: Can Iteration Drive Innovation?

Evgeny Kagan
Carey Business School, Johns Hopkins University, ekagan@jhu.edu

Tobias Lieberum
, tobias.lieberum@tum.de

Sebastian Schiffels
University of Augsburg, sebastian.schiffels@uni-a.de

Christian Jost
Technical University Munich, christian.jost@tum.edu

Problem definition: Motivated by the widespread adoption of Agile, Scrum and other iterative project management techniques, we study the effects of workflow – iterative or sequential – on innovative behavior and performance. ***Methodology/Results:*** We conduct a series of laboratory experiments. Our first experiment shows that, in an open-ended creative challenge, iterative task completion leads to better outcomes than sequential task completion. In the second experiment we show that the advantage of iterative workflow further extends to innovation settings that do not involve idea generation. A key mechanism driving the advantage of iterative work is that it leads to frequent task switching, prompting workers to perform a broader search for the best available solution. In the third experiment we delve deeper into the search process and show that sequential work indeed leads to more myopic idea refinement behaviors, often ending in a (sub-optimal) local maximum. ***Managerial implications:*** Our results suggest that iterative workflow improves performance across multiple, structurally distinct, innovation settings. We also identify three boundary conditions. First, iterative workflow helps achieve quick gains, but its performance advantage narrows over time. Second, iterative workflow mainly helps low performers but has minimal effects on top performers. Third, iterative workflow can be harmful in projects with strong path dependencies between subtasks.

1. Introduction

Effective time management is central to most innovation activities. The way workers navigate tasks – sequentially, focusing on one task at a time or *iteratively*, working on multiple tasks concurrently and completing them in increments rather than whole – can greatly affect efficiency, the quality of output, and the potential to generate and develop ideas. In this paper we report the results of a series of experiments examining the effects of sequential and iterative workflow on innovative behavior and performance.

Our research is motivated by the widespread adoption of new product development paradigms that emphasize iterative progression of innovation activities. One such paradigm is *Agile* – a suite of project management techniques characterized by shorter iterative cycles, continuous feedback, and worker autonomy (Laufer et al. 2015, Rigby et al. 2016, Kettunen and Lejeune 2020, Allon et al. 2021, Lieberum et al. 2022, Ghosh and Wu 2023). A central element of an Agile process is a "product backlog" – a dynamic, prioritized collection of project features, enhancements, and bug fixes that outlines the individual workflow for each person working on a project. The backlog is continuously updated and reprioritized to align with evolving project goals, customer needs and technological advancements. The fluidity, while beneficial for adapting to change, often leads to frequent changes in focus as workers are required to constantly reorient their efforts, transitioning between different tasks and features, and completing tasks in smaller increments to enable frequent prototyping. In contrast, the more traditional *Waterfall* development model promotes sequential task completion, wherein each task is completed before work on the next one may begin.

Our interactions with Allianz, a large multinational insurance company, offer insightful perspectives on how an organizational shift from Waterfall to Agile affects workflow and task management. As is common for many organizations, Agile was first rolled out in the IT function before spreading to other functions, including marketing and sales. For the purposes of this paper we restrict our observations to a specific unit within Allianz IT function, whose primary responsibility is the development and maintenance of an internal software platform used by Allianz' employees to perform various actuarial tasks. The following insights emerged from our interactions:

- **Two-week sprints and product backlogs:** Allianz' shift to Agile has led to the adoption of two-week *sprints* – well-defined, time-boxed intervals, during which each developer is assigned a set of items from a product backlog. Each sprint concludes with a prototype demo, offering feedback on each developer's incremental performance and ending with selection, prioritization and assignment of new tasks for the subsequent sprint.
- **Iterative workflow leads to more task fragmentation and task switching:** The shift to Agile has led to a significant increase in task fragmentation, both at the project and at the individual developer level. Allianz' developers now face a more diverse array of tasks in

each sprint and work on each task in shorter increments. The prioritization of backlog items, driven by their strategic importance (rather than task type), has led to a departure from linear workflows. The development process has become more iterative, with features undergoing development, testing, and launch in accelerated and overlapping cycles.

- **Increased autonomy:** A key part of the transition is the empowerment of individual developers through increased autonomy. While all tasks should be completed by the end of the sprint, developers are encouraged to exercise their judgment and independently navigate between tasks and features as the situation demands. For example, developers are asked to actively recognize possibilities for technological improvements and adjust priorities accordingly.

The sum of these observations suggests that the Agile model can lead to substantial changes in individual workflow – it can rearrange the workflow into ever smaller increments, increase task switching, and lead to greater worker autonomy. How such transformations affect innovation performance is not immediately clear. On the one hand, worker autonomy can improve performance through greater process ownership and can be an enabler of creativity (Hackman and Oldham 1976, Zhou 1998, Sawyer 2011, Wuttke et al. 2022). Being able to switch between tasks can help workers identify how their efforts can be spent most productively. At the same time, the removal of constraints and the need to concurrently balance multiple priorities and make a larger number of decisions can also harm performance through increased setup costs and cognitive overload (Salvucci and Taatgen 2008, Lurie and Swaminathan 2009, Kagan et al. 2018, Long et al. 2020, Colicev et al. 2023).

To examine the effects of iterative vs sequential workflow on innovative behavior and performance we conduct a series of laboratory experiments, in which we carefully vary the workflow (iterative vs. sequential) in several, structurally distinct, innovation settings. Our experimental design draws on the experimental psychology tradition (Guilford 1950, Torrance 1966, Simonton 2000, Sawyer 2011), which focuses largely on idea generation and brainstorming stages of innovation activities. Additionally, we leverage the innovation/search literature in economics and management (Levinthal and March 1981, Levinthal 1997, Mihm et al. 2003, Ederer and Manso 2013, Billinger et al. 2014, Sommer et al. 2020), which focuses on the more downstream stages of innovation. Our experiments examine behaviors related to both the generative stages where ideas are created, as well as behaviors related to the later stages of innovation where the best ideas are selected and combined into a single integrated whole.

Our first experiment is an open-ended creative challenge. The experimental activity is a variation on the popular game "Scrabble" – participants are given letters of the alphabet and must build connected verbal structures under time and material constraints. The Scrabble activity consists of two tasks: one task is a Scrabble board where participants can only use verbs, while the second task is a board where they can only use nouns. In the sequential workflow treatment participants first complete one task (e.g., nouns), and then move on to the second one (e.g., verbs). In contrast, with an iterative workflow participants complete a full iteration of the activity, splitting their time into shorter increments as they work on both tasks (nouns and verbs), repeatedly switching between tasks as they see fit.

Our main experimental result is that iterative workflow outperforms sequential, with performance improvements ranging between 15 and 28 percentage points. In additional treatments we rule out the explanation that iterative workflow allows workers to spend more time on the more difficult or value-adding task. Indeed, participants in all treatments spend approximately the same amount of time on each task. That is, improved productivity is caused not by better overall time allocation, but by frequent task switching and refocusing one's attention on a new problem. Examining participant productivity, we find that the low performance of the sequential treatment is driven mainly by sharply diminishing marginal gains as workers approach the end of the time period allocated to each task. This is especially true during the initial work period, suggesting that participants run out of ideas faster with sequential workflow, especially when the task is new to them and they are required to generate new ideas in an unfamiliar environment. In contrast, iterative workflow appears to stimulate creative production at a more steady pace.

The Scrabble activity includes both a creative idea generation component (forming words), and a more integrative, combinatorial component (connecting new words with existing ones). To examine whether the advantage of iterative work extends to settings that lack the idea generation component, we conduct a second experiment in which participants are supplied with a set of ideas rather than having to generate new ideas. Similar to the first experimental activity this activity is based on Scrabble. However, different from the classic Scrabble task participants now receive a list of *pre-formed* words, and the task is to use as many of these words as possible to build a single connected structure. Consistent with our earlier results, we find that in this setting iterative workflow continues to outperform sequential, suggesting that the performance gap is not driven solely by differences in idea generation and creative production, but also extends to more downstream innovation activities.

A key performance indicator in the second experiment is the number of restarts of the search for new solutions (participants removing words from the board and starting from scratch). Indeed, such restarts occur more frequently with iterative workflow, suggesting that iterative workflow stimulates more exploratory behaviors, while sequential workflow leads to more narrow and myopic strategies. To better understand this mechanism, we introduce a third experiment in which we use a version of the multi-dimensional search task ("Lemonade Stand Task", Ederer and Manso 2013, Sommer et al. 2020), and zoom into explore-exploit behaviors as a potential driver of the advantages of iterative work. Comparing search behavior and performance across different parametrizations of the search landscape, we find that iterative workflow is beneficial on more "rugged" (or interdependent) landscapes. In contrast, sequential workflow performs just as well as iterative on smoother landscapes where the "greedy", myopic hill-climbing approach is a viable strategy.

In addition to search landscape "ruggedness", we are able to identify three boundary conditions on the benefits of iterative workflow. First, iterative workflow helps low performers but does not affect top performance. Second, iterative workflow leads to quicker gains than sequential, but the performance gap becomes more narrow over time. Third, iterative workflow may reduce performance when subtasks have strong path dependencies. If the work completed in the early stages cannot be easily altered, iterative workflow may hinder the exploration of alternatives and lock the worker into a suboptimal path.

2. Literature and Contributions

Innovation and creativity research spans various fields, including psychology, economics, strategy, and operations management (Krishnan and Ulrich 2001, Sørensen et al. 2010, Sawyer 2011, Kavadias and Ulrich 2020). However, there remains a significant gap in our understanding of how workflow affects innovation behavior and performance. No study, to our knowledge, examines the micro-level dynamics of task management in innovation-related tasks. We next discuss three streams of literature that inform our experiment design, and that our study contributes to: the task selection literature, the project management literature, as well as the broader innovation and creativity research that uses real-effort tasks.

2.1. Task Ordering and Selection

The study of people-centric operations, i.e., of how human factors influence operational processes, has gained attention in recent years (Roels and Staats 2021). A notable stream within this literature is research on task ordering and selection, conducted mainly in medical settings. Ibanez et al. (2018) study physicians' task prioritization and observe a tendency to choose the shortest or easiest tasks first. This finding is further validated by Kc et al. (2020), who show that such behaviors negatively affect performance in both lab and field. If these patterns extend to our innovation setting, we should expect suboptimal performance with the more flexible, iterative workflow, as workers may allocate more of their time to less complex tasks, rather than to those contributing most value. To ensure that workers do not overspend time on easier or more enjoyable task components, one of our iterative workflow treatments restricts the time spent on each task to be the same. Kc and Terwiesch (2011) use hospital (and hospital unit) data and show positive effects of focus on performance. Similarly, Staats and Gino (2012) find that task variety can have short-term negative effects, which would also speak for sequential workflow in our setting. Kc (2014) finds that some task switching can enhance productivity and quality of care, but also cautions against its overuse.

Finally, Siemsen (2008) and Katok and Siemsen (2011) examine task selection in principal-agent settings where workers use the difficulty level of tasks to signal ability and garner greater rewards. Our experiments abstract away from any interactions between workers and managers, and instead use individual tasks with objectively measurable performance.

2.2. Project Management

The closest related project management studies are Kagan et al. (2018) and Lieberum et al. (2022). Kagan et al. (2018) find that designers who decide for themselves how to spend time between creative ideation and execution perform worse than designers with exogenously imposed schedules – an effect driven mainly by delays in worker-determined schedules (see also, Goldratt 1990, Ariely and Wertenbroch 2002). Relatedly, Bendoly et al. (2014) find that managerial progress checks are key to effective task switching. Lieberum et al. (2022) show that time-boxing of work, i.e., imposing fixed time intervals for tasks, can improve performance. They use a slider task, which measures pure effort (as opposed to innovation performance). While these studies examine work arrangements that give workers more/less process control and autonomy, none of them compare iterative vs. sequential workflow, or explore multiple, structurally distinct, innovation settings.

2.3. Experimental Tasks in Innovation Literature

Two of our experimental activities build on the experimental psychology tradition of using verbal tasks to study creative behaviors (Sawyer 2011). A common approach is to use verbal puzzles or riddles (Kachelmeier et al. 2008, Kachelmeier and Williamson 2010, Erat and Gneezy 2016) or deciphering anagrams (Mendelsohn and Griswold 1964, Ansburg and Hill 2003, O'Connor et al. 2013). Many of these tasks emphasize the creative ideation component of the innovation process, where the objective is to produce as many ideas as possible. One of our experimental activities is based on the popular game "Scrabble". Similar to the literature, this game also requires participants to generate many ideas; however, in addition to idea generation it also reproduces the more downstream stages of innovation, such as the need to select, combine and implement the best ideas into a single integrated product. The creative energy is directed towards a pragmatic, performance-oriented goal, as participants seek to build the largest possible structure under tight time and material constraints.

Our third experimental activity leverages the approach (more common among economists and business disciplines) of representing innovation as a multidimensional search process (Levinthal and March 1981, Levinthal 1997, Mihm et al. 2003, Sommer and Loch 2004). In this setting, idea generation is muted; instead, each potential solution is a vector of product attributes, and the worker's goal is to identify the best combination among a very large number of possibilities, typically under time constraints. To achieve good performance the worker needs to develop an understanding of the mapping between combinations of product attributes and the resulting performance. While the theoretical literature on complex solution landscapes is exhaustive (in particular for NK models; see Baumann et al. 2019, for a recent review), the number of experiments examining human search strategies on a landscape is relatively small. See Ederer and Manso (2013) and Sommer et al. (2020) for recent examples. A key advantage of the Lemonade Stand task is that it allows the experimenter to manipulate landscape "ruggedness" – a key moderator of the effects of workflow on performance in our setting.

2.4. Contributions

Our study is the first systematic effort that we are aware of, to explore how individual workflow influences innovation behaviors and outcomes across multiple, structurally distinct innovation settings. We contribute to the task selection and project management literature by providing a novel test of key project management techniques within an innovation setting. Our research also contributes to the creativity literature in experimental psychology and management, which has primarily studied individual features of innovators, such as their personality and team dynamics, and mainly relies on idea generation tasks (Sawyer 2011). By examining multiple, distinct stages of the innovation process we offer a more comprehensive perspective on the factors driving innovation performance.

Table 1 Experiment Overview

Workflow (Treatments varied between-subject)	**Questions and activities** (Each subject completes two activities: a version of the Scrabble game and a version of the Lemonade Stand game)	
	§4: How does workflow affect performance in a setting that has both a creative and a combinatorial component? **Activity: Scrabble**	***§6: Do the treatment effects replicate in a different innovation setting?*** **Activity: Lemonade Stand (rugged landscape)**
T1: Sequential	*SEQ*	*SEQ*
T2: Iterative, time and process constraints	*ITER EQUAL FREEZE*	*ITER EQUAL FREEZE*
T3: Iterative, process constraints	*ITER FREEZE*	*ITER FREEZE*
T4: Iterative, no constraints	*ITER*	*ITER*
	§5: How does workflow affect performance in a setting that has only the combinatorial component? **Activity: Scrabble with pre-formed words**	***§6 (cont'd): Does myopic, "greedy" optimizing drive the disadvantage of sequential workflow?*** **Activity: Lemonade Stand (smooth landscape)**
T5: Sequential	*SEQ*	*SEQ*
T6: Iterative, no constraints	*ITER*	*ITER*

Notes. *ITER* stands for iterative workflow, *SEQ* stands for sequential workflow. *EQUAL* means that the total amount of time allocated to task 1 must be equal to the time allocated to task 2. *FREEZE* means that the choices made in period 1 cannot be altered in period 2. Ordering of activities (Scrabble ⟶ Lemonade Stand or Lemonade Stand ⟶ Scrabble) was randomized in the experiment.

3. Experiment overview

To examine the effects of workflow on innovative behavior and performance we conducted a series of laboratory experiments. The experiments were organized into a 6 (*treatments*, between-subject) × 2 (*activities*, within-subject) design. Table 1 summarizes the treatments and activities.

3.1. Activities and Treatments

In each treatment subjects completed two activities, administered in random sequence: a version of Scrabble, and a version of the Lemonade Stand game. The *Scrabble activity* requires participants to engage in both creative idea generation (forming words), and idea recombination and integration (combining words in a performance-maximizing manner). We examine this activity in §4. The *Scrabble with pre-formed words activity* removes the creative idea generation element but retains the recombination element. We examine this activity in §5. The *Lemonade Stand activity* is an experimental game designed specifically to study innovative behaviors related to idea recombination (Ederer and Manso 2013, Sommer et al. 2020). The advantage of this activity is that it allows a more tightly controlled test of treatment effects on explore-exploit behaviors. In particular, we examine a more rugged landscape where broader exploration is needed to achieve good results, and a smoother landscape where more narrow, myopic search can be sufficient. We examine this activity in §6.

Within each activity subjects worked on two tasks. In particular, the Scrabble activity included a Scrabble board for verbs only, and a Scrabble board for nouns only. Analogously, the Lemonade Stand activity included two separate search landscapes, one related to market attributes of the lemonade stand, and a second one related to product attributes. Depending on the treatment, subjects were either required to complete the tasks in a pre-determined sequence (*SEQ*), or were allowed to switch between tasks, thereby completing the tasks iteratively (*ITER*). In addition, we examined two intermediate regimes, in which subjects worked iteratively, but were not allowed to alter the work performed in the first period (*ITER FREEZE*), as well as a regime in which they worked iteratively, but were required to spend equal amounts of time in each task (*ITER EQUAL FREEZE*). Both of these treatments impose constraints on the iterative workflow, bringing it closer to a sequential workflow and allowing us to identify some of the key mechanisms.

3.2. Payments and Protocols

All experiments were conducted at a large public German University. The experiments were programmed and conducted in German, the first language of most of the participants. Sessions consisted of ten to twelve participants. The sequence of activities and the sequence of tasks within each activity were randomized to control for order effects and fatigue. Participants were paid a fixed show-up fee of EUR 5 and a variable payment based on their performance in each of the two real-effort tasks. The average total payment was EUR 11.33. The total duration of the experiment was 45 minutes, resulting in average hourly earnings of EUR 15.11. (The laboratory target earnings rate was EUR 14/hour at the time of data collection.) The experimental interface was programmed in o-Tree (Chen et al. 2016). A total of 479 participants were recruited across all six treatments. Participants were only admitted to each activity after passing several attention and comprehension checks. See EC.1-EC.2 for the full protocol, instructions and exclusions.

4. Scrabble

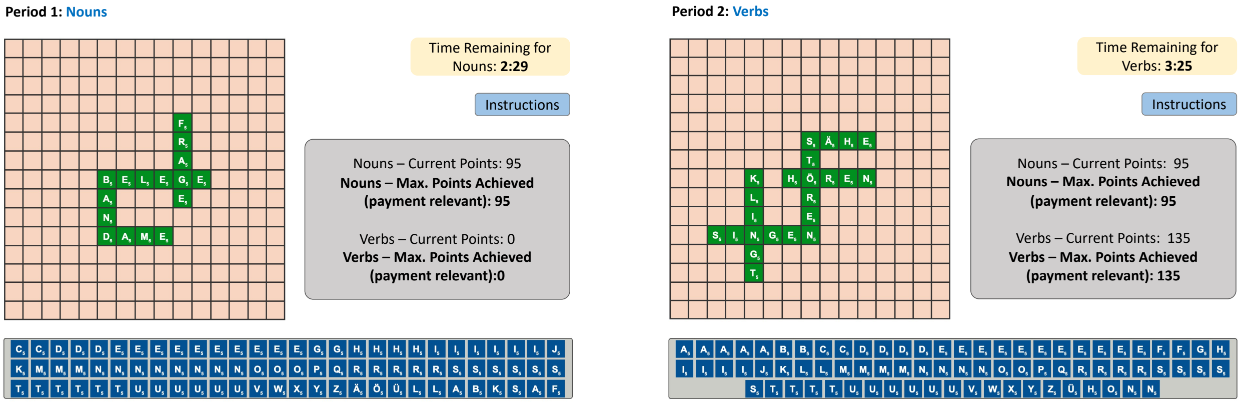
In this section we examine the effects of workflow in a Scrabble-based activity. The open-ended, creative nature of Scrabble lends itself to studying the relevant behaviors and drivers of performance in a setting where the solution space is very large and ex ante unknown, and participants must discover, explore and recombine various ideas to achieve good results.

4.1. Experimental Setup and Hypotheses

4.1.1. Setup At the start of the Scrabble activity, participants receive a set of tiles with letters on them. Words must be formed and connected in crossword fashion, and must read left to right or top to bottom. Deviating from the classic version of Scrabble, there are two separate boards that represent two product features or components. On one board subjects may only form nouns, and on the other board they may only form verbs. Each board has 15 × 15 fields. For each board, subjects receive 100 letters with no refill. The list of letters is the same for each participant. Words cannot be formed diagonally. Sample screen shots are in Figure 1. The validity of each word placed on the board is instantly checked against the online dictionary wiktionary.org and highlighted in green color if valid. The overall performance, used to determine participant compensation, was computed as the minimum of the two task scores (verb and noun scores). This is to represent that a product has multiple components, and each of the components needs to be done well before the

Figure 1 **Scrabble: Screenshots**

(a) Task 1: Nouns (b) Task 2: Verbs



Period 1: Nouns

Time Remaining for Nouns: **2:29**

Nouns – Current Points: 95
Nouns – Max. Points Achieved (payment relevant): 95

Verbs – Current Points: 0
Verbs – Max. Points Achieved (payment relevant):0

Period 2: Verbs

Time Remaining for Verbs: **3:25**

Nouns – Current Points: 95
Nouns – Max. Points Achieved (payment relevant): 95

Verbs – Current Points: 135
Verbs – Max. Points Achieved (payment relevant): 135

Instructions
FRAGE
BELEGE
BAND
DAME
C C D D D E E E E E E E E E E G G H H H H I I I I I I J
K M M M N N N N N N N N O O O P Q R R R R R S S S S S S
T T T T T T U U U U U U V W X Y Z Ä Ö Ü L L A B K S A F
Instructions
SÄHE
STÖREN
HÖREN
KLINGT
SINGEN
A A A A A B B C C D D D D E E E E E E E E E E E E F F G H
I I I I J K L L M M M M N N N N O O P Q R R R R S S S S
S T T T T U U U U U U V W X Y Z Ü H O N N

Note: Sample screen shots (translated from German) for the sequential workflow treatment.

Table 2 Scrabble Treatments

Treatment	Work periods	Task switching allowed within period?	Time allocation to tasks flexible?	Can period 1 work be altered during period 2?
SEQ	Two periods, six minutes each	No	No	No
ITER EQUAL FREEZE	Two periods, six minutes each	Yes	No	No
ITER FREEZE	Two periods, six minutes each	Yes	Yes	No
ITER	Two periods, six minutes each	Yes	Yes	Yes

product can be taken to market. This payoff function also ensures that participants are incentivized to work on both tasks, instead of working on the task they consider to be easier or more enjoyable. Each used letter was worth five points.

4.1.2. Treatments (Between-Subject) We consider four between-subject treatments. In all four treatments the overall time is 12 minutes, and there are two periods of equal length. The workflow, i.e., the allocation of time to tasks depends on treatment. In the *SEQ* treatment participants complete the activity sequentially, with only one task being worked on in each period. In the *ITER* treatment participants work on both task in parallel throughout the time horizon, switching between tasks as they see fit. In the *ITER EQUAL FREEZE* treatment, the words placed on the board in the first period are "frozen" during the second period and cannot be (re-)moved. In addition, in this treatment participants must allocate exactly the same amount of time to each task. Thus, the only difference between *SEQ* and *ITER EQUAL FREEZE* is that participants in the latter treatment are allowed to task-switch. The *ITER FREEZE* treatment is analogous, but the equal time allocation constraint is relaxed. The differences in workflow between the treatments are summarized in Table 2.

4.1.3. Hypotheses The standard economic argument is that a less constrained action set should improve performance. In our setting, *SEQ* presents workers with the most constraints, while *ITER* is the least-constrained workflow (See Table 1). We use this reasoning to formulate the following hypothesis regarding the effects of workflow on performance.

H1: *Treatment performance is ranked as follows: SEQ < ITER EQUAL FREEZE < ITER FREEZE < ITER.*

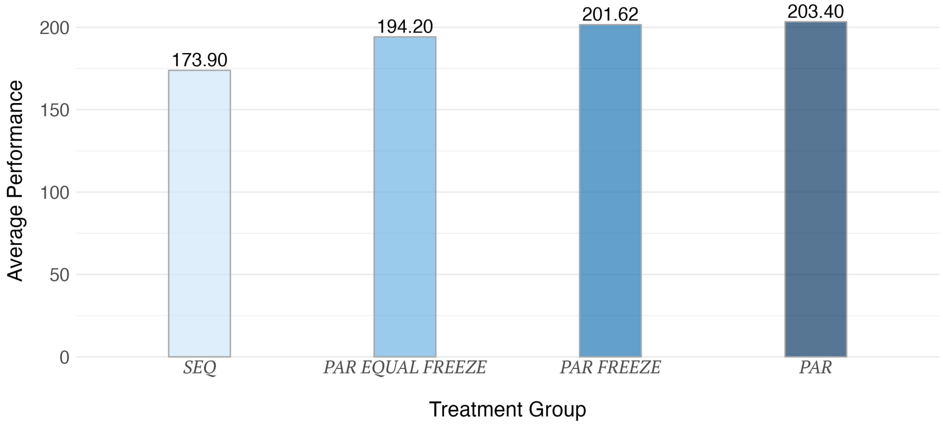
Counterarguments to H1 are found in several behavioral studies showing that constraints can be helpful in some complex tasks (Lurie and Swaminathan 2009, Sawyer 2011, Kagan et al. 2018, Long et al. 2020). Other research found that when given a choice, workers overspend time and energy on the easier, rather than the most value-adding tasks (Ibanez et al. 2018, Kc et al. 2020, see §2.1 for details). While it is not clear whether these types of behaviors will occur in our setting, these studies suggest that certain types of constraints may indeed be beneficial for performance.

4.2. Results

4.2.1. Summary Statistics Figure 2 shows average treatment performance. The lowest performance is observed in the sequential workflow treatment (*SEQ*). Further, the largest gap is between the *SEQ* and the iterative treatments. Both of these patterns are in line with H1: fewer workflow constraints lead to better performance. However, some of the treatment differences are minimal. In particular, while the differences in mean performance range from 20.3 to 29.5 points between *SEQ* and iterative treatments, the differences within the iterative treatments are minimal (ranging between 1.8 and 9.8 points). This suggests that some of the H1 predictions may not be supported in the data, and that the key driver of performance differences is the ability to task-switch, rather than the presence of additional time and process constraints.

4.2.2. Hypothesis Tests We next test H1 using regression analysis. Table 3 shows the estimates of the treatment effects, with the *SEQ* treatment indicator serving as the baseline for comparisons. Columns (1)-(2) show the effects on overall performance (participant payoff) - with and without demographic control variables. Both specifications show robust effects of all iterative treatments on performance, ranging from 24.41 to 46.39 points (between 14.75 and 27.72 percentage point improvement, computed based on marginal effects), with four of the six $p-$values below 0.01, and two of the $p-$values equal to 0.044 and 0.060, respectively. Examining the performance in more detail in columns (3)-(4), we find that the performance gap between *SEQ* and the iterative

Figure 2 **Scrabble: Performance by Treatment**



treatments is driven primarily by the poor performance of sequential workflow in the first task, i.e., during the first work period (column 3). In contrast, if we focus on the second task (column 4), the differences between treatments are smaller and not statistically significant. Finally, the bottom panel of Table 3 shows that the differences among the iterative treatments are minimal. Taken together, these analyses suggest that the ability to switch back-and-forth and multitask is a key performance differentiator between iterative and sequential workflows. In contrast, time and process constraints play a subordinate role.

4.2.3. Detailed Analysis We next discuss three sets of analysis that help unpack the drivers of treatment effects: the differences in the performance distributions within each treatment (heterogeneous treatment effects), the differences in the types of ideas (words) used in each treatment, as well as the differences in productivity and timing.

Heterogeneous treatment effects The positive effects of iterative workflow are strongest in the left tail of the performance distribution. For example, the 25th percentile of performance is 110 points in *SEQ*, and ranges between 145 and 180 points in the iterative treatments. In contrast, the 75th percentiles are quite similar across the four treatments, ranging between 233 and 255. That is, iterative workflow primarily helps low performers, but does not meaningfully affect high performers. As a result, performance variance in each iterative treatment is lower relative to *SEQ*. Indeed, Levene's test of equality of variance rejects the null at $p \ll 0.01$ for all pairwise comparisons

Table 3 Scrabble: Hypothesis Tests

	(1)	(2)	(3)	(4)
Dependent Variable:	*Performance*	*Performance*	*First task score*	*Second task score*
SEQ	Baseline	Baseline	Baseline	Baseline
ITER EQUAL FREEZE	46.39^{***}	46.19^{***}	47.68^{***}	18.53
	(14.28)	(13.94)	(14.81)	(13.92)
ITER FREEZE	43.06^{***}	40.70^{***}	35.94^{***}	3.577
	(12.66)	(12.29)	(13.06)	(12.27)
ITER	26.78^{**}	24.41^{*}	32.03^{**}	-10.52
	(13.22)	(12.93)	(13.75)	(12.91)
Constant	164.5^{***}	161.7^{***}	154.6^{***}	227.7^{***}
	(10.86)	(28.32)	(30.10)	(28.28)
Demographic controls	No	Yes	Yes	Yes
Observations	244	244	244	244
R-squared	0.104	0.182	0.189	0.166
Pairwise tests ($p-$values)				
ITER EQUAL FREEZE = ITER FREEZE	0.804	0.675	0.399	0.253
ITER EQUAL FREEZE = ITER	0.217	0.162	0.343	0.062
ITER FREEZE = ITER	0.252	0.240	0.790	0.309

Notes. OLS regressions with standard errors in parentheses. Total score is the lower of the two task scores. All specifications control for task sequence, task sequence, loss of internet connection. Columns (2)-(4) control for age, gender, German native speaker, education, and familiarity with Scrabble. $^{*}p < 0.10$; $^{**}p < 0.05$; $^{***}p < 0.01$.

between *SEQ* and each iterative treatment. Detailed analysis of treatment effect heterogeneity is presented in Appendix EC.3 (Quantile regressions in Table EC.3.4).

Qualitative differences in the words used There were few qualitative differences in the types of words formed in each treatment. In particular, we found that the average length of words (and word extensions) was quite similar between treatments (between 5.24 and 5.55, pairwise rank sum tests $p > 0.301$). Further, we found no significant differences in the uniqueness of the words used by participants, with all pairwise $p > 0.132$. (Uniqueness metric is the average coincidence score between the words used by a participant against all words used by the subject pool). That is, neither of the treatments prompted participants to use words that were significantly more complex or particularly unusual. Finally, we examined whether participants reuse words with the same root across tasks (verbs derived from nouns or vice versa), and found that this did not explain the advantage of iterative treatments (none of the treatment differences were significant, with $p > 0.100$).

Number of words and timing While the length and nature of words were quite similar between treatments, there were some significant differences in the number of words formed, particularly in the first period. In particular, participants in *SEQ* formed an average of 7.35 words during the first period, while participants in the iterative treatments formed between 8.46 and 9.32, depending on the treatment (pairwise rank sum test between *SEQ* and each iterative treatment: $p = 0.018, p = 0.100, p = 0.001$). Further, examining minute-by-minute changes in performance, in *SEQ* the number of words declined sharply over time within each period (non-parametric trend test $p = 0.000$ in both periods). In contrast, the number of words remained close to constant over time in each of the iterative treatments (all $p > 0.359$). To better understand the benefits of task switching, we also examined the timing of words placed in the iterative treatments right before and right after each task switch. The amount of time between the last placed word and a task switch was 29 seconds. The amount of time *after* a task switch was 27 seconds. Notably, the latter includes the time needed to move multiple tiles to the board, suggesting quite high levels of productivity immediately following a switch.

Together these dynamics highlight that sequential workflow leads to sharply diminishing gains over time, particularly in the first work period. In contrast, task switching appears to be an effective strategy that enables participants in the iterative workflow to maintain productivity at a more constant pace.

5. Scrabble with Pre-formed Words

The purpose our next experiment is twofold. First, the Scrabble activity in §4 combines creative and analytical thinking, with good performance relying on successful idea generation, selection, and execution. This complicates the separate observation and measurement of idea formation and integration with existing pieces, as many ideas are discarded internally before they manifest on the board. Second, in practice, innovation often happens in the later stages of development which emphasize integrating and recombining existing ideas rather than creating new ideas from scratch. For example, Agile teams at Allianz, discussed in §1, mainly utilize existing code and designs to build new products. To broaden our insights, in this section we will introduce a new experimental activity that helps isolate these more downstream innovation behaviors from idea generation.

5.1. Experimental Setup and Hypotheses

The general experimental protocol (participants, payments, exclusions etc.) is described in §3. Here we focus on the key changes in the design relative to the Scrabble activity of §4.

5.1.1. New Scrabble Activity The new activity retains the physical/spatial design component of the Scrabble activity requiring participants to combine elements in a 2D space, but restricts them from producing new ideas. Instead, participants need to examine different combinations of existing ideas. To this end, we use a smaller, 5×5 Scrabble board and require participants to *only* use combinations of 20 five-letter pre-formed words, either nouns or verbs (We made the board smaller relative to §4 to ensure that the task was sufficiently difficult). As before, the goal is to use as many letters as possible. We chose a single combination of words leading to a global optimum (six words, i.e., 30 letters, equivalent to 150 points) and two combinations that lead to the next best result (125 points). Screenshots are in Figure EC.1.1 in the Appendix.

We administered two between-subject treatments: *SEQ* and *ITER*. Similar to the previous version of the Scrabble activity, in the *SEQ* treatment participants were only allowed to work on a single task (verbs or nouns) in each of the two periods. However, in the *ITER* treatment they were allowed to switch between tasks and work iteratively. Due to the lack of notable differences among the different iterative workflow treatments in §4, we omit the intermediate *ITER FREEZE* and *ITER FREEZE EQUAL* treatments in this activity.

5.1.2. Hypotheses The open-ended, creative nature of the Scrabble activity means that there are several potential pathways for the treatment effects identified in §4. One possible pathway is that iterative workflow facilitates idea production and removes creative blockages - a performance barrier that has been documented in the idea generation and brainstorming literature (Sawyer 2011, and references there). Working concurrently on two tasks (verbs and nouns) may benefit creative thinking and help unblock creative production, leading to the generation of a greater number of ideas. If this mechanism is the main driver of the treatment effects, then the performance advantage of iterative workflow should collapse once idea generation is muted. Furthermore, the disadvantages of iterative work, such as the cognitive costs of switching between tasks (Gilovich et al. 2002, Colicev et al. 2023), as well as potential increases in the cognitive load (Lurie and Swaminathan 2009, Kagan et al. 2018, Long et al. 2020, see §2.1-2.2 for details) may neutralize any benefits of increased flexibility. This leads us to the following hypothesis:

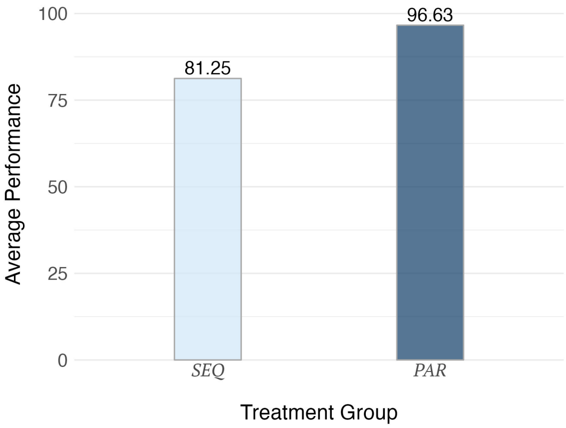
H2: *When participants are provided with pre-existing ideas (words), average performance does not differ between SEQ and ITER.*

An alternative pathway for the benefits of iterative workflow is that it does not affect the generative stages of idea production, but rather the selection and integration of ideas. Girotra et al. (2010) and Kagan et al. (2018) highlight the importance of idea selection in creative performance, showing that certain time management strategies prevent inertia and improve selection quality. If this holds in our setting, H2 would be rejected and iterative workflow would continue to dominate.

5.2. Results

5.2.1. Summary Statistics Figure 3 shows average performance by treatment. Similar to the classic Scrabble activity used in §4, there is a notable difference in treatment performance. Further, the magnitude of the difference is quite similar to the original Scrabble activity: the improvement going from *SEQ* to *ITER* is 18.93%. This provides preliminary evidence counter to H2, i.e., even when idea generation is muted, iterative workflow continues to outperform sequential.

Figure 3 Scrabble with Pre-formed Words: Performance by Treatment



5.2.2. Hypothesis Tests To test H2 more formally, we conduct a series of regressions, analogous to the tests in §4.2.2. The regression coefficients are reported in Table 4. The regression results do not support H2: *ITER* workflow continues to dominate both in the absence of demographic controls ($p = 0.013$, col. 1), and after controlling for individual differences ($p = 0.019$, column 2). As before, the treatment difference is statistically significant only if we compare performance in the first displayed task ($p = 0.010$, column 3), but not in the second task ($p = 0.797$, column 4). Taken together, these results replicate the performance patterns observed in §4, suggesting that idea generation cannot be the sole driver of our results. Instead, the ability to recognize effective combinations on a complex solution space of ideas appears be an important pathway through which iterative workflow improves performance.

5.2.3. Detailed Analysis A key behavior characterizing the participant's approach is the number of times the participant removes a word from the board. This is because the action space is constrained by the small size of the board, so that performance improvements are only possible through repeated removal and repositioning of words. Examining the frequency of word additions and word removals, we find that on average, both are significantly higher with iterative flow (Additions: 9.73 times in *ITER* vs. 11.81 times in *SEQ*, $p = 0.001$; Removals: 3.48 times in *ITER* vs. 2.26 times in *SEQ*, $p = 0.024$). Further, the frequency of both word additions and word removals

Table 4 Scrabble with Pre-formed Words: Hypothesis Tests

	(1)	(2)	(3)	(4)
Dependent Variable:	*Performance*	*Performance*	*First task score*	*Second task score*
SEQ	Baseline	Baseline	Baseline	Baseline
ITER	15.15**	13.65**	14.85**	1.233
	(5.981)	(5.712)	(5.664)	(4.790)
Constant	76.95***	125.8***	136.2***	120.1***
	(5.588)	(16.35)	(16.21)	(13.71)
Demographic controls	No	Yes	Yes	Yes
Observations	112	112	112	112
R-squared	0.082	0.234	0.246	0.130

Notes. OLS regressions with standard errors in parentheses. All specifications control for task sequence, task sequence, loss of internet connection. Columns (2)-(4) control for age, gender, German native speaker, education, and familiarity with Scrabble. $^{*}p < 0.10$; $^{**}p < 0.05$; $^{***}p < 0.01$.

was strongly correlated with performance (Pearson correlation coefficient for additions and performance: $\rho = 0.71$, $p \ll 0.001$; for removals: $\rho = 0.33$, $p = 0.001$). Finally, the heterogeneous treatment effects identified in §4 persist: despite improving performance average, *ITER* leads to a substantially smaller performance variance (*SEQ*: 1260.25, *ITER*: 650.25; Levene's test $p = 0.073$).

Together, these comparisons indicate that the advantage of iterative workflow extends to settings in which ideas do not need to be generated, but rather need to be recombined and integrated in a performance-maximizing manner. Iterating appears to prompt participants to go back to the "drawing board" and restart the search for better idea combinations, while sequential workflow appears to lead to more linear, myopic behaviors. In the next section, we will delve deeper into the search and selection processes, focusing in particular on explore-exploit behaviors and will examine more systematically how these behaviors affect innovation performance.

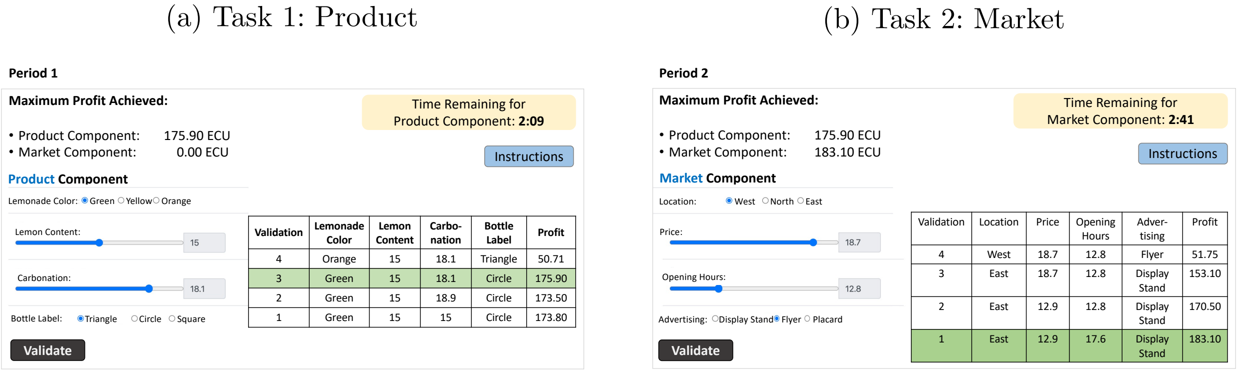
6. Lemonade Stand

We have so far explored the effects of iterative and sequential workflow using relatively unstructured tests of creative behavior. Although such tasks can help increase external validity (compared to more structured tests of decision-making), they offer limited control and insight into the decision processes and behaviors involved in navigating the solution landscape. In this section, we introduce a new experimental activity (Lemonade Stand game, Ederer and Manso 2013), which allows more precise control over both the solution landscape and the behaviors that drive innovative performance. This activity will serve as a robustness test for our prior results, and will help us identify several key boundary conditions on the benefits of iterative workflow.

6.1. Experimental Setup and Hypotheses

6.1.1. Lemonade Stand Activity The approach of representing innovation activities as search on multidimensional landscapes is standard in the economics and management research (Levinthal and March 1981, Levinthal 1997, Mihm et al. 2003, Sommer and Loch 2004, Billinger et al. 2014). As is common in the experimental implementation of landscape problems (see, for example, Ederer and Manso 2013, Sommer et al. 2020) we use the naturalistic framing of designing

Figure 4 Lemonade Stand Activity: Screenshots



(a) Task 1: Product

Period 1

Maximum Profit Achieved:

- Product Component: 175.90 ECU
- Market Component: 0.00 ECU

Time Remaining for Product Component: **2:09**

Product Component

Lemonade Color: ● Green ○ Yellow ○ Orange

Lemon Content: 15

Carbonation: 18.1

Bottle Label: ● Triangle ○ Circle ○ Square

Validation	Lemonade Color	Lemon Content	Carbo-nation	Bottle Label	Profit
4	Orange	15	18.1	Triangle	50.71
3	Green	15	18.1	Circle	175.90
2	Green	15	18.9	Circle	173.50
1	Green	15	15	Circle	173.80

(b) Task 2: Market

Period 2

Maximum Profit Achieved:

- Product Component: 175.90 ECU
- Market Component: 183.10 ECU

Time Remaining for Market Component: **2:41**

Market Component

Location: ● West ○ North ○ East

Price: 18.7

Opening Hours: 12.8

Advertising: ○ Display Stand ● Flyer ○ Placard

Validation	Location	Price	Opening Hours	Adver-tising	Profit
4	West	18.7	12.8	Flyer	51.75
3	East	18.7	12.8	Display Stand	153.10
2	East	12.9	12.8	Display Stand	170.50
1	East	12.9	17.6	Display Stand	183.10

Note: Sample screenshots (translated from German) for the *SEQ* treatment. As in previous work using the lemonade stand game (Ederer and Manso 2013, Sommer et al. 2020), we use a mix of discrete and continuous attributes. Specifically, lemonade color, bottle label, location and advertising are discrete attributes. The remaining attributes are continuous. The continuous attributes allow inputs in the $[10, 20]$ range, with the choices limited to one digit after the decimal point, yielding a total of 101 possible choices each. Thus, the solution space in each task has $3 \times 3 \times 101 \times 101 = 92,000$ unique combinations. The tables show each examined combination, with the best discovered combination highlighted in green.

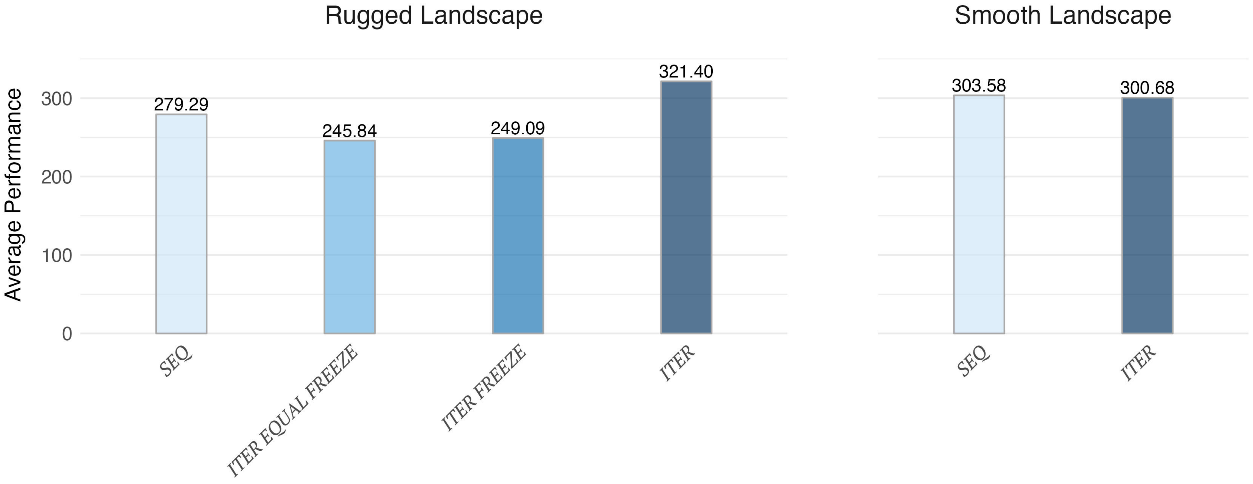
and managing a ”Lemonade stand” to represent the solution landscape. The participant is asked to identify an effective business strategy by repeatedly choosing the values of several business attributes, and learning about the payoff resulting from each attribute combination. Deviating from the classic version of the Lemonade Stand game, we introduce two separate independent landscapes: one for product and one for market attributes. The product landscape consists of four product attributes: lemonade color, lemon content, carbonation, bottle label. The market landscape also consists of four market attributes: location, price, opening hours, advertising. Figure 4 shows the decision screens for each of the two tasks. Participants can modify the attributes as often as they like. However, each time they do so, there is a 3 second delay until they see the resulting profit. This is to encourage thoughtful choices and to discourage random clicking. As before, participants were paid based on the lower of the two task scores.

6.1.2. Treatments We examined two different versions (parametrizations) of the Lemonade Stand game: a *rugged* landscape parametrization, and a *smooth* one. These parametrizations were chosen and calibrated based on previous implementations of the Lemonade Stand game (for example, Ederer and Manso 2013, Sommer et al. 2020). The landscapes are visualized in Figures EC.2.2-EC.2.5. In all parametrizations there is a single global optimum with a payoff of 500 points, and two further local optima, earning them 200 and 380 points, respectively. In the rugged parametrization, discovering the global optimum requires more experimentation. This is because the locations of each of the three optima are different in each of the nine combinations of the discrete attributes. In contrast, in the rugged parametrization, "greedy" myopic refinement can suffice to discover the global optimum. This is because the attributes are less interdependent, i.e., each attribute has its own optimal value that does not depend on other attributes. For example, in the smooth parametrization, it is always optimal to choose a price of 17.9 units, regardless of the other attributes. In contrast, in the rugged parametrization, the optimal price depends on the choice of lemonade color and bottle label.

As noted in §3 (Table 1), we conduct four treatments in the rugged landscape parametrization, and only two of the treatments (*SEQ* and *ITER*) in the smooth parametrization. This is because the effects of freezing are minimal in the Scrabble game and do not call for extensive replication. To impose freezing constraints in the Lemonade Stand game (*ITER EQUAL FREEZE* and *ITER FREEZE* treatments), we fix two of the four attributes in each task to their best discovered values after the first period. Thus, participants in these two treatments are somewhat more limited in their actions during the second phase.

6.1.3. Hypotheses To develop hypotheses we return to the theoretical discussions in §2.1-2.2 and in §4.1.3. Having more flexibility to choose how to allocate time and being able to iterate should generally improve performance. Based on our previous results, the hypothesized pathway through which iterative workflow may facilitate performance improvements is repeated task switching. Task switching splits each task into smaller increments, with each increment presenting an opportunity for a fresh start. In the context of the Lemonade Stand game, such opportunities are especially useful when the solution landscape is more rugged – in this case restarting the search may prompt participants to experiment with more diverse combinations of attributes and can thus help perform a broader exploration of the landscape. In contrast, workflow should have minimal effects on smoother landscapes, where myopic, "greedy" optimizing of individual features can lead to good results. We formalize this logic as follows:

Figure 5 **Lemonade Stand: Performance by Treatment**



H3A: *On a rugged landscape, average performance is ranked as follows: SEQ < ITER EQUAL FREEZE < ITER FREEZE < ITER.*

H3B: *On a smooth landscape, average performance does not differ between SEQ and ITER.*

6.2. Results

6.2.1. Summary Statistics Figure 5 shows average performance in each treatment group. The figure suggests several marked differences between treatments. First, the left part of the figure shows that *ITER* outperforms *SEQ* in the rugged landscape parametrization, with a performance improvement of 42.11 points. Second, the performance in both *ITER EQUAL FREEZE* and *ITER FREEZE* is lower relative to *ITER* and also relative to *SEQ*, suggesting that the effects of freezing constraints are more substantial, compared to the Scrabble activity. However, as before, the equal time allocation constraint does not appear to affect performance: the difference between *ITER EQUAL FREEZE* and *ITER FREEZE* is only 3.25 points. Finally, the right part of the figure suggests only minimal effects of workflow in the smooth landscape parametrization (2.90 point difference between treatments).

6.2.2. Hypothesis Tests Table 5 shows regression coefficients in the rugged (columns 1-4) and smooth (columns 5-8) landscape versions of the Lemonade Stand game. The results confirm that most of the treatment differences in performance observed in the left part of Figure 5 (rugged

Table 5 Lemonade Stand Activity: Regression Results

	Rugged landscape				Smooth landscape			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Dep. Var.	*Performance*	*Performance*	*First task score*	*Second task score*	*Performance*	*Performance*	*First task score*	*Second task score*
SEQ	Baseline	Baseline	Baseline	Baseline	Baseline	Baseline	Baseline	Baseline
ITER EQUAL FREEZE	-31.67** (14.98)	-32.15** (15.13)	9.058 (17.81)	-69.13*** (17.35)				
ITER FREEZE	-29.39* (14.99)	-29.70* (15.09)	6.853 (17.77)	-71.77*** (17.31)				
ITER	44.09*** (16.87)	44.36** (17.16)	54.01*** (20.20)	-16.29 (19.68)	-0.515 (17.91)	-5.448 (18.42)	0.623 (19.95)	-27.26 (20.53)
Constant	280.6*** (14.26)	283.0*** (36.76)	292.9*** (43.28)	401.0*** (42.17)	297.5*** (20.22)	385.8*** (52.68)	383.3*** (57.06)	462.8*** (58.73)
Demographic controls	No	Yes	Yes	Yes	No	Yes	Yes	Yes
Observations	286	286	286	286	115	115	115	115
R-squared	0.090	0.093	0.042	0.147	0.036	0.071	0.037	0.070

Notes. OLS regressions with standard errors in parentheses. Demographic controls are age, gender, German native speaker, education. $^{*}p < 0.10$; $^{**}p < 0.05$; $^{***}p < 0.01$.

landscape) are significant. In particular, the treatment effect of *ITER* is 44.09 points, corresponding to an improvement of 15.7 percentage points relative to *SEQ*. The comparison is significant at $p = 0.009$. Further, both *ITER EQUAL FREEZE* and *ITER FREEZE* are worse than *SEQ* ($p = 0.034$ and $p = 0.050$), confirming that the freezing constraint is binding in this case. Restricting the actions in the second phase significantly reduces performance. Further, column (3) shows that the benefits of iterative workflow, again, are caused by the poor performance of *SEQ* in the first phase, with the treatment effect of *ITER* being quite large (54.01 points) and significant at $p = 0.008$. In contrast, the effect dissipates in the second phase ($p = 0.409$ in column 4). Finally, columns (5)-(8) confirm that none of the treatment differences are statistically significant in the smooth landscape parametrization (all $p > 0.187$).

6.2.3. Detailed Analysis We discuss two sets of additional analyses. First, as with the previous activities, there were some distributional differences in performance. In the Lemonade Stand game, performance is typically clustered in the vicinity of local optima (Ederer and Manso 2013). In Table 6 we summarize the share of participants reaching each optimality region. The comparisons show that iterative workflow again mainly helps low performers: comparing *SEQ* and *ITER*, the

Table 6 Lemonade Stand Activity: % of Subjects in Each Region

	Rugged Parametrization				**Smooth Parametrization**	
	SEQ	*ITER EQUAL FREEZE*	*ITER FREEZE*	*ITER*	*SEQ*	*ITER*
% of subjects reaching **low** local optimum region	52.2%	65.8%	62.0%	20.0%	35.0%	25.5%
% of subjects reaching **middle** local optimum region	28.3%	24.7%	28.2%	56.0%	46.7%	60.0%
% of subjects reaching **global** optimum region	19.6%	9.6 %	9.9%	24.0%	18.3%	14.5%
Total	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%

proportion of participants stuck in the bottom local optimum decreases from 52.2% to 20.0% in the rugged parametrization and from 35.0% to 25.5% in the smooth parametrization (Proportion tests, both $p < 0.01$). Further, freezing constraints prevent the majority of participants in *ITER FREEZE* and *ITER EQUAL FREEZE* (62 and 66%, respectively) from escaping the lowest region. Comparing performance variances, we again find that the variances in all iterative treatments are smaller than in *SEQ* (with $p-$values between 0.002 and 0.080 depending on the treatment).

Second, we examine several performance drivers, focusing in particular on the number of explored solutions and on the concentration/dispersion of these solutions across the solution space. We find that the total number of attempted solutions does not explain the treatment differences; in fact, the average number of attempted solutions is greater in *SEQ* than in *ITER* (52 vs. 48). However, the concentration of solutions is an important predictor of performance. Greater concentration (measured by the Herfindahl-Hirschman Index, averaged across all attributes) is negatively correlated with performance in the rugged parametrization ($\rho = 0.196, p < 0.01$), but not in the smooth parametrization ($\rho = 0.084, p = 0.370$). This is because the smooth landscape can be searched effectively without broad exploration. We also find that participants in *ITER* perform a more dispersed search than those in *ITER*, particularly in the first period. These comparisons suggest that sequential workflow leads to more myopic, “greedy” hill-climbing search behaviors leading to poor performance on more rugged landscapes. Conversely, iterative workflow facilitates a more dispersed search for the best solution. Therefore, landscape ruggedness serves as an important moderator of the benefits of iterative workflow.

Table 7 Summary of Results

	Activity	Nature of task	Preferred workflow	Boundary conditions
§4	Scrabble	Idea generation & recombination	Iterative	1) Significant productivity differences in first period, not in second; 2) Iterative workflow helps bottom performers, not top performers.
§5	Scrabble with pre-formed words	Recombination	Iterative	
§6	Lemonade stand, rugged landscape	Recombination	Iterative	1) Significant productivity differences in first period, not in second; 2) Iterative workflow helps bottom performers, not top performers; 3) Iterative workflow can be harmful in projects with strong path dependencies between subtasks (freezing).
§6	Lemonade stand, smooth landscape	Recombination	No significant differences	

Finally, participants in *ITER FREEZE* and *ITER EQUAL FREEZE* are forced to be even more concentrated in their search, because portions of the solution are unavailable to them in the second period, due to freezing. Further analysis shows that this is because they fail to anticipate the freezing of attributes, and are thus locked into a suboptimal path during the second period.

7. Integrated Discussion, Managerial Implications and Conclusions

The results of our studies are summarized in Table 7. Our main result is that iterative workflow consistently outperforms sequential workflow in multiple, structurally different innovation environments. As shown in Table 7, iterative workflow led to higher average performance in three out of four cases. The benefits are statistically significant and economically meaningful – switching to the iterative approach resulted in average marginal performance gains of up to 28 percent.

We also identified three boundary conditions. First, the advantage of iterative workflow was statistically significant in the initial work phase, but was minimal in the second work phase. Second, different workflow treatments led to different performance distributions. The iterative approach mainly helped low performers, with top performers largely unaffected. This was true for all treatments comparisons in our study, regardless of the experimental activity. Finally, we saw that iterative workflow can harm performance when there are strong path dependencies between subtasks. When later iterations included constraints on the action space, many workers found themselves stuck on suboptimal exploration paths, unable to discover the global optimum region.

While the ability to iterate and switch between tasks helped productivity, we found that this was not explained by improved time allocation. Indeed, in the Scrabble activity, performance was the same whether the time spent in each task was exogenously imposed to be equal (*ITER EQUAL FREEZE* treatment) or endogenously determined by the worker (*ITER* and *ITER FREEZE* treatments). This (null) result is surprising given that a less constrained action set should improve productivity. However, this result is consistent with the growing body of work that finds that more autonomy may not always improve performance in complex tasks, such as product design (Kagan et al. 2018), project selection and abandonment (Long et al. 2020), and time and effort allocation (Lieberum et al. 2022). We contribute to this literature, clarifying that while certain types of autonomy, such as the ability to switch between tasks, can be helpful, other types, such as flexible time allocation, may not.

The result that iterative workflow improves average, though not necessarily top performance has meaningful implications for firms that manage a portfolio of innovation projects. Our results suggest that when the objective is to avoid failure or to maximize average performance iterative workflow is often preferred. However, if the objective is to maximize top performance and failure can be tolerated, then the sequential approach may still be viable.

Further unpacking the performance differences, we found sequential workflow to underperform during the first, but not necessarily during the second phase. Thus, while there was evidence of improvement in sequential treatments – indicating some learning – this progress was not enough to offset the initial gap. An implication of this result is that when iterative workflow cannot be implemented, organizations may benefit from giving workers the discretion to arrange the order of tasks as they see fit. In particular, our learning-related results suggest that workers may perform better if they tackle the easiest or most familiar tasks first. Innovation activities may therefore present an exception to more routine work settings, where completing easy tasks first has been shown to reduce performance (Ibanez et al. 2018, Kc et al. 2020).

Notably, the reduced initial performance observed with sequential workflow is different from a “cold-start” effect, i.e., a delay in the initial build – a behavior observed under some conditions in creative tasks (Kagan et al. 2018). In contrast, in our setting productivity gains were quite similar across treatments at the outset, but slowed down around the middle of the first working period when working sequentially. Our subsequent experiments showed that the key mechanism driving this was the myopic search for narrow improvements, and the lack of a broader exploration of the solution space.

While most complex tasks involve some learning, creative blocks and slowdowns are a unique feature of creative processes (Sawyer 2011). The finding that such performance barriers can be mitigated through iterative work is a novel result that is useful for developing more accurate theories of innovative behavior. This result also provides some empirical grounding for future models of entrepreneurial and innovation activities, for example when linking effort, time and performance in innovation tournaments (Terwiesch and Xu 2008), especially where workers divide their time and attention between multiple simultaneous assignments (Körpeoğlu et al. 2022, Kızılyıldırım et al. 2022), or for models of entrepreneurial time allocation (Yoo et al. 2016).

Opportunities for creating a more iterative workflow exist in many organizational settings. In product development, moving from the sequential waterfall model to a more iterative, agile approach allows workers to task-switch more frequently, as they make multiple, small adjustments to multiple features in each iteration, rather than devoting their full attention to one feature at a time. Early-stage startups, typically operating under tight time-to-market constraints, can also benefit from the iterative approach. For example, they can benefit from exploring multiple different markets or customer segments in parallel, rather than conducting market research sequentially, one segment at a time.

The experimental tasks used in our study represent only a subset of all innovation settings. It is therefore important to highlight several key limitations of our experiments. First, we have focused on settings in which tasks are independent (i.e., do not require integration) and are only linked through a payoff function. Thus, our findings may not apply in settings where performance is driven to a large extent by component integration or assembly, for example in aerospace R&D or other complex engineering settings. Future work may expand on our experiment by adding a third phase where participants must integrate completed components. Second, our study focused on individual activities with precise and immediate performance feedback. Interpersonal dynamics, both in collaborative teams found in R&D, and in competitive contexts like innovation tournaments, might offer interesting variations. Finally, settings where the tasks are very different and where task switching would require greater setup costs present further potential boundary conditions that can be explored.

References

Allon G, Askalidis G, Berry R, Immorlica N, Moon K, Singh A (2021) When to be agile: Ratings and version updates in mobile apps. *Management Science* Forthcoming:1–19.

Ansburg PI, Hill K (2003) Creative and analytic thinkers differ in their use of attentional resources. *Personality and Individual Differences* 34(7):1141–1152.

Ariely D, Wertenbroch K (2002) Procrastination, deadlines, and performance: Self-control by precommitment. *Psychological science* 13(3):219–224.

Baumann O, Schmidt J, Stieglitz N (2019) Effective search in rugged performance landscapes: A review and outlook. *Journal of Management* 45(1):285–318.

Bendoly E, Swink M, Simpson III WP (2014) Prioritizing and monitoring concurrent project work: Effects on switching behavior. *Production and Operations Management* 23(5):847–860.

Billinger S, Stieglitz N, Schumacher TR (2014) Search on rugged landscapes: An experimental study. *Organization Science* 25(1):93–108.

Chen DL, Schonger M, Wickens C (2016) oTree—An open-source platform for laboratory, online, and field experiments. *Journal of Behavioral and Experimental Finance* 9:88–97.

Colicev A, Hakkarainen T, Pedersen T (2023) Multi-project work and project performance: Friends or foes? *Strategic Management Journal* 44(2):610–636.

Ederer F, Manso G (2013) Is pay for performance detrimental to innovation? *Management Science* 59(7):1496–1513.

Erat S, Gneezy U (2016) Incentives for creativity. *Experimental Economics* 19(2):269–280.

Ghosh S, Wu A (2023) Iterative coordination and innovation: Prioritizing value over novelty. *Organization Science* 34(6):2182–2206.

Gilovich T, Griffin D, Kahneman D (2002) *Heuristics and biases: The psychology of intuitive judgment* (Cambridge university press).

Girotra K, Terwiesch C, Ulrich KT (2010) Idea generation and the quality of the best idea. *Management science* 56(4):591–605.

Goldratt EM (1990) *Theory of constraints* (North River Croton-on-Hudson).

Greiner B, et al. (2004) The online recruitment system orsee 2.0-a guide for the organization of experiments in economics. Technical report.

Guilford J (1950) Creativity. *American Psychologist* 5(9):444–454.

Hackman JR, Oldham GR (1976) Motivation through the design of work: Test of a theory. *Organizational Behavior and Human Performance* 16(2):250–279.

Ibanez MR, Clark JR, Huckman RS, Staats BR (2018) Discretionary task ordering: Queue management in radiological services. *Management Science* 64(9):4389–4407.

Kachelmeier SJ, Reichert BE, Williamson MG (2008) Measuring and motivating quantity, creativity, or both. *Journal of Accounting Research* 46(2):341–373.

Kachelmeier SJ, Williamson MG (2010) Attracting creativity: The initial and aggregate effects of contract selection on creativity-weighted productivity. *The Accounting Review* 85(5):1669–1691.

Kagan E, Leider S, Lovejoy WS (2018) Ideation–execution transition in product development: An experimental analysis. *Management Science* 64(5):2238–2262.

Katok E, Siemsen E (2011) Why genius leads to adversity: Experimental evidence on the reputational effects of task difficulty choices. *Management Science* 57(6):1042–1054.

Kavadias S, Ulrich KT (2020) Innovation and new product development: Reflections and insights from the research published in the first 20 years of manufacturing & service operations management. *Manufacturing & Service Operations Management* 22(1):84–92.

Kc DS (2014) Does multitasking improve performance? evidence from the emergency department. *Manufacturing & Service Operations Management* 16(2):168–183.

Kc DS, Staats BR, Kouchaki M, Gino F (2020) Task selection and workload: A focus on completing easy tasks hurts performance. *Management Science* 66(10):4397–4416.

Kc DS, Terwiesch C (2011) The effects of focus on performance: Evidence from california hospitals. *Management Science* 57(11):1897–1912.

Kettunen J, Lejeune MA (2020) Waterfall and agile product development approaches: Disjunctive stochastic programming formulations. *Operations Research* 68(5):1356–1363.

Kızılyıldırım R, Korpeoglu CG, Körpeoglu E, Kremer M (2022) Exclusive or not? an experimental analysis of parallel innovation contests.

Körpeoğlu E, Korpeoglu CG, Hafalır İE (2022) Parallel innovation contests. *Operations Research* 70(3):1506–1530.

Krishnan V, Ulrich KT (2001) Product development decisions: A review of the literature. *Management science* 47(1):1–21.

Laufer A, Hoffman EJ, Russell JS, Cameron WS (2015) What Successful Project Managers Do. *MIT Sloan Management Review* 43–51.

Levinthal DA (1997) Adaptation on rugged landscapes. *Management science* 43(7):934–950.

Levinthal DA, March JG (1981) A model of adaptive organizational search. *Journal of Economic Behavior & Organization* 2(4):307–333.

Lieberum T, Schiffels S, Kolisch R (2022) Should we all work in sprints? How agile project management improves performance. *Manufacturing & Service Operations Management* Forthcoming:1–17.

Long X, Nasiry J, Wu Y (2020) A behavioral study on abandonment decisions in multistage projects. *Management Science* 66(5):1999–2016.

Lurie NH, Swaminathan JM (2009) Is timely information always better? the effect of feedback frequency on decision making. *Organizational Behavior and Human decisión processes* 108(2):315–329.

Mendelsohn GA, Griswold BB (1964) Differential use of incidental stimuli in problem solving as a function of creativity. *The Journal of Abnormal and Social Psychology* 68(4):431–436.

Mihm J, Loch C, Huchzermeier A (2003) Problem–solving oscillations in complex engineering projects. *Management Science* 49(6):733–750.

O'Connor AJ, Nemeth CJ, Akutsu S (2013) Consequences of beliefs about the malleability of creativity. *Creativity Research Journal* 25(2):155–162.

Rigby DK, Sutherland J, Takeuchi H (2016) Embracing agile: How to master the process that's transforming management. *Harvard Business Review* 94(5):40–50.

Roels G, Staats BR (2021) Om forum—people-centric operations: Achievements and future research directions. *Manufacturing & Service Operations Management* 23(4):745–757.

Salvucci DD, Taatgen NA (2008) Threaded cognition: an integrated theory of concurrent multitasking. *Psychological review* 115(1):101.

Sawyer RK (2011) *Explaining creativity: The science of human innovation* (Oxford university press).

Siemsen E (2008) The hidden perils of career concerns in r&d organizations. *Management Science* 54(5):863–877.

Simonton DK (2000) Creativity: Cognitive, personal, developmental, and social aspects. *American psychologist* 55(1):151.

Sommer SC, Bendoly E, Kavadias S (2020) How do you search for the best alternative? Experimental evidence on search strategies to solve complex problems. *Management Science* 66(3):1395–1420.

Sommer SC, Loch CH (2004) Selectionism and learning in projects with complexity and unforeseeable uncertainty. *Management Science* 50(10):1334–1347.

Sørensen F, Mattsson J, Sundbo J (2010) Experimental methods in innovation research. *Research policy* 39(3):313–322.

Staats BR, Gino F (2012) Specialization and variety in repetitive tasks: Evidence from a japanese bank. *Management science* 58(6):1141–1159.

Terwiesch C, Xu Y (2008) Innovation contests, open innovation, and multiagent problem solving. *Management science* 54(9):1529–1543.

Torrance EP (1966) Torrance tests of creative thinking. *Educational and Psychological Measurement* .

Wuttke D, Upadhyay A, Siemsen E, Wuttke-Linnemann A (2022) Seeing the bigger picture? ramping up production with the use of augmented reality. *Manufacturing & Service Operations Management* 24(4):2349–2366.

Yoo OS, Corbett CJ, Roels G (2016) Optimal time allocation for process improvement for growth-focused entrepreneurs. *Manufacturing & Service Operations Management* 18(3):361–375.

Zhou J (1998) Feedback valence, feedback style, task autonomy, and achievement orientation: Interactive effects on creative performance. *Journal of applied psychology* 83(2):261.

Electronic Companion

EC.1. Experimental Protocol and Instructions

Experiments were programmed in o-Tree (Chen et al. 2016). Participants were recruited via ORSEE (Greiner et al. 2004). A total of 479 participants were recruited. A total of 38 participants were not admitted the experiment because they had failed the German test. A total of 2 participants were excluded due to technical issues. A total of 83 participants were not admitted to the Scrabble activity because they failed the attendant comprehension test, and a total of 38 participants were not admitted to the Lemonade Stand activity because they failed the attendant comprehension test. If a participant was excluded from one of the two activities, they did not receive the payoff from that activity, and continued to the second activity.

Due to Covid-19 restrictions, all experiments were conducted online. Zoom was used to monitoring the participants. Zoom meetings were set up with at least one of the authors as a hosts. Participants received Zoom links via email in the morning of the day of the experiment. Upon sign-up, participants were renamed to preserve anonymity. During the experiment participants were able to chat with the experimenter and ask questions. All instructions were read loud. The instructions are summarized below (translated from German):

Introduction

Welcome to today's experiment. The experiment will take about 45 minutes. Participation in the experiment is only possible with the Google Chrome browser and a computer mouse. Participation with another browser as well as with cell phone or tablet is not possible due to technical reasons. If you do not meet this condition, you cannot participate in the experiment. In this case, please leave the Zoom meeting now.

Please leave your camera on for the entire duration of the experiment. This is only to ensure that everything runs smoothly. There will be no recording. By voluntarily participating in this experiment, you expressly consent to this use in accordance with the General Data Protection

Regulation. If you do not want to agree to the camera use, you can leave the Zoom meeting now without further consequences. If you lose your Internet connection during processing, dial into this Zoom meeting again. We will then explain the further procedure to you.

Do you have any questions? Then write a private message to the lead experimenter via the Zoom chat. There are several comprehension tests. Do not hesitate to write to me if something is unclear.

We will now send you a custom link through Zoom chat. Copy and paste it into your Chrome browser. You can start working on it right away. When you reach the end of the experiment, you can leave this Zoom meeting and close the experiment.

Thank you for participating in this scientific study!

Opening Screens

Welcome to today's experiment! It's good to have you with us!

This is an individual experiment. To ensure scientific validity, the tasks vary between the participants of this experiment. Therefore, please do not attempt to interact with each other or third parties. The use of cell phones, tablets, software, and internet applications other than this experiment is strictly prohibited for the entire duration of the experiment. Violations will result in exclusion from further participation in experiments in the [lab name blinded for review]. Do not press the reload, back, or forward buttons on your browser, or the F5 key, as this will cancel the experiment. Please keep your camera turned on throughout the experiment. If you have any questions, please write us a private message to the experimenter in Zoom Chat.

As announced in the invitation of the experiment, a confident command of the German language is important for this experiment. Therefore, you must first pass a German test.

[Followed by the German test.]

Part 1 of the Experiment - Instructions and Comprehension Test

[Note: part 1 and part 2 of the experiment were displayed in random order.]

Please read the following instructions carefully and answer the comprehension questions. You will have two attempts to pass the comprehension questions. If you do not successfully pass the *comprehension questions, you will not participate in this part of the experiment and will not be compensated for it. If you have any questions about the instructions, please write a PRIVATE message to the experiment director using the Zoom chat function.*

Background

In this part of the experiment, you will develop the most profitable business model for a lemonade stand by selecting a product and market strategy from numerous options. The product task consists of four product characteristics:

1. *Color*
2. *Lemon content*
3. *Carbon dioxide content*
4. *Bottle label*

The market task consists of four market characteristics:

1. *Location*
2. *Price*
3. *Opening hours*
4. *Advertising*

On the computer screen you can choose different combinations of the product and market characteristics. For this purpose, you can change single, several or all characteristics of a component at the same time. Then click on the "Validate selection" button to see the profit resulting from your selection. This is displayed in the fictitious currency ECU. In a table you can see all your combinations validated so far and their profitability.

Within a component, all characteristics influence the profitability. However, your decisions on the product strategy do not influence the profitability of the market strategy and vice versa.

The most profitable combination in each case has been defined by chance. Therefore, do not try to draw conclusions about the best strategy from your own experience outside the experiment, but explore the respective circumstances without bias. For example, do not let your life experience *guide you as to which lemon content or price customers would value most, but test the taste and willingness to pay in the experiment. Please note that product and market components are equally important for the success of your business model, i.e. the maximum achievable profit each from product and market strategy is identical.*

Your task

There are two game phases during which you can develop your strategies. Both phases last four minutes each. In between you have a break of 30 seconds.

[*SEQ treatment:*] *During the first phase, you can work exclusively on the product strategy; during the second phase, you can work exclusively on the market strategy. You can change and validate the characteristics as many times as you want within a phase. However, your decisions in the first phase (the four characteristics color, lemon content, carbon dioxide content, and bottle label for product strategy) are set and cannot be changed during the second phase.*

[*ITER treatments:*] *During both phases, you are free to decide how long you work on the product and market strategy. To do this, you can switch back and forth between the two components. You can change and validate the characteristics as many times as you want within a phase. However, four of the eight characteristics (color and lemon content for product strategy, location and price for market strategy) are set after the first phase based on the highest profit achieved and cannot then be changed during the second phase.*

Your compensation

[All treatments:] *Your compensation depends on the profitability of each of your product and market strategies. First, the combination with the highest profit is selected separately for each product and market component from all trials. That is, it is not the last chosen combination that is decisive, but the most profitable one. Second, for your business to be successful, both product and market components must convince customers. Thus, you will be paid the LOWER profit from product and market strategy.*

The following example illustrates the payoff (the profit values shown are arbitrarily chosen and not representative). You have tried five combinations for your product strategy and three combinations for your market strategy:

Table EC.1.1

Product strategy	Profit
Combination 1	ECU 20
Combination 2	ECU 10
Combination 3	ECU 60
Combination 4	ECU 30
Combination 5	ECU 20
Market strategy	**Profit**
Combination 1	ECU 50
Combination 2	ECU 30
Combination 3	ECU 10

First, the combination with the highest profit is determined for product and market strategy individually. In our example, this is combination 3 for the product strategy and combination 1 for the market strategy. Second, you are paid the lower profit of the two strategies, i.e. in this case, Combination 1 of the market strategy (ECU 50). The higher profit of the product strategy (ECU 60) is not paid out. The exchange rate is ECU 70 = EUR 1.00.

EC.1.1. Part 2 of the Experiment

Please read the following instructions carefully and answer the comprehension questions. You will have two attempts to pass the comprehension questions. If you do not successfully pass the comprehension questions, you will not participate in this part of the experiment and will not be compensated for it. If you have any questions about the instructions, please write a PRIVATE message to the experiment director using the Zoom chat function.

Background

In this part of the experiment, you will form German nouns and verbs (no adjectives, names, brands, cities, etc.) from letters, each on its own playing field, similar to Scrabble. Declension and conjugation forms are allowed. There are 100 different letters available for each game field.

You must place the first letter on the orange square in the middle of the game field. Further letters must always be placed directly on other letters and cannot be placed without this connection.

All letter combinations must make valid words from left to right and top to bottom, but not diagonally. A word is only valid if it is listed at Wiktionary.org (Wiktionary.org is a word collection similar to the Duden). It is then displayed in green.

Your task

There are two phases of the game during which you can form words. Both phases last six minutes each. In between you have 30 seconds break.

[*SEQ:*] *During the first phase, you can work exclusively on the playfield for nouns; during the second phase, you can work exclusively on the playfield for verbs.*

[*ITER:*] *During both phases, you are free to decide how long you work on the game board for nouns and the game board for verbs, respectively. To do this, you can switch back and forth between the two playing fields indefinitely.*

[All treatments:] *Letters can be changed and removed only during the phase in which they are placed, i.e. letters that you have placed in the first phase cannot be changed or removed in the second phase. To remove letters, drag them from the edge of the letter field back into the letter pool.*

Your compensation

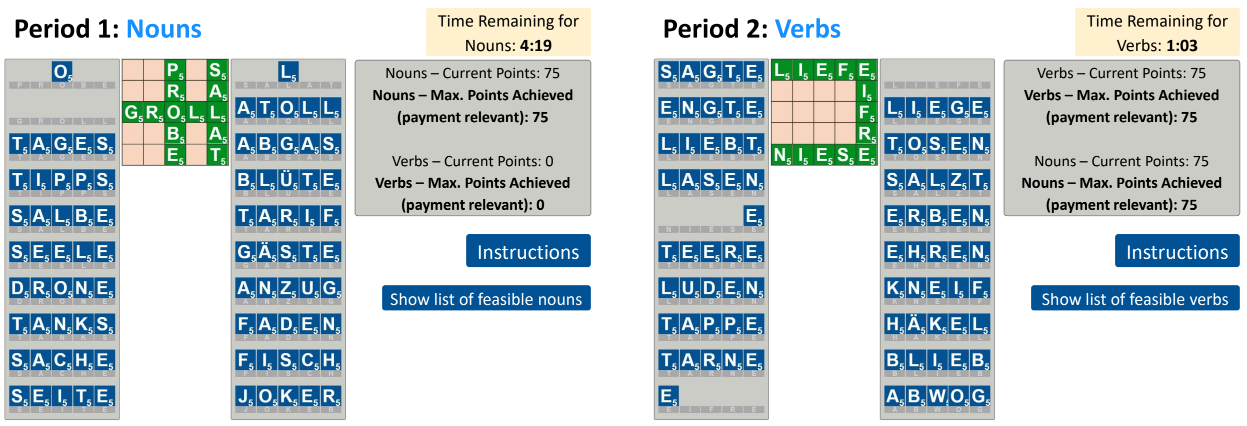
Your compensation depends on the number of correctly placed letters on both playing fields.

*First, the correctly placed letters are counted separately for each of the two game fields. Letters used for two words are counted twice. Each letter is worth 5 points. There are no bonus points, each word is counted only once and each valid letter gives the same score. For example, if there are 2 words with 4 and 6 letters on one board, the score is (4+6) * 5 = 50 points.*

If not all placed letters result in valid words, the game field is invalid and the highest score before the game field became invalid is valid. Therefore, the current score can be lower than the highest score. For example, if you fail to finish a word in the last seconds of the editing time, the highest score before you started the invalid word counts.

You will be paid only the LOWER of the score of both fields. For example, if you have accumulated 50 points for nouns and 60 points for verbs, you will be paid 50 points (these point values are arbitrarily chosen and are not representative). The exchange rate is 70 points = EUR 1.00.

Figure EC.1.1 Scrabble with Pre-formed Words: Screenshots



(a) Task 1: Nouns

Period 1: Nouns

Time Remaining for Nouns: **4:19**

Nouns – Current Points: 75
Nouns – Max. Points Achieved (payment relevant): 75

Verbs – Current Points: 0
Verbs – Max. Points Achieved (payment relevant): 0

(b) Task 2: Verbs

Period 2: Verbs

Time Remaining for Verbs: **1:03**

Verbs – Current Points: 75
Verbs – Max. Points Achieved (payment relevant): 75

Nouns – Current Points: 75
Nouns – Max. Points Achieved (payment relevant): 75

EC.1.2. Scrabble with Pre-formed Words: Screenshots

Figure EC.1.1 shows screenshots of the interface of the Scrabble with pre-formed words activity (discussed in §5). Similar to the original Scrabble activity, this activity also consisted of two tasks: one board with nouns and a second board with verbs only. Each participant received the same list of 20 nouns and 20 verbs, displayed in Figure EC.1.1. As before each letter placed on the board was worth 5 points. There was a single "global " maximum (placing a total of six words on the board, resulting in a payoff of 150 points) and two suboptimal, "local" maxima (placing a total of five words, resulting in a payoff of 125 points). The remaining experimental details (task durations, random ordering etc.) were unchanged relative to the original Scrabble activity of §4.

EC.2. Experimental Design Details and Parametrization

In this section we describe the implementation details of all three activities, in particular, the materials provided to participants during the Scrabble activities, and the parametrization of the Lemonade Stand activity.

EC.2.1. Scrabble Activity (§4)

Following the classic German version of Scrabble, 100 tiles were made available to the subjects for each (Noun and Verb) task. The tiles were not refilled for the second period. The tiles given to participants at the beginning of the task were as follows (number of tiles with each letter is given in parentheses):

E (15), N (9), S (7), I (6), R (6), T (6), U (6), A (5), D (4) H (4), G (3), L (3), O (3) M (4), B (2), W (1), Z (1) C (2), F (2), K (2), P (1) Ä (1), J (1), Ü (1), V (1) Ö (1), X (1) Q (1), Y (1)

EC.2.2. Scrabble Activity with Pre-formed Words (§5)

See Figure EC.1.1 for the complete list of words for this activity.

EC.2.3. Lemonade Stand Activity (§6)

We developed an adaptation of the classic Lemonade Stand game (Ederer and Manso 2013), which includes two separate tasks, each with a separate, independent solution landscape. The first task is the Product component, consisting of four attributes (lemonade color, lemon content, carbonation, shape of the bottle label). The second component is the Market component, consisting of four attributes (location, price, opening hours, advertising). For each task two of the attributes are discrete, with three levels to choose from, while the other two are continuous, and can be varied in increments of 0.1 units. Figure EC.2.2 and Figure EC.2.3 show the landscapes for all combinations of the discrete variables for the product and the market tasks, for the *Rugged* parametrization.

Subjects were presented with two tasks (Product and Market), with each component containing four parameters.

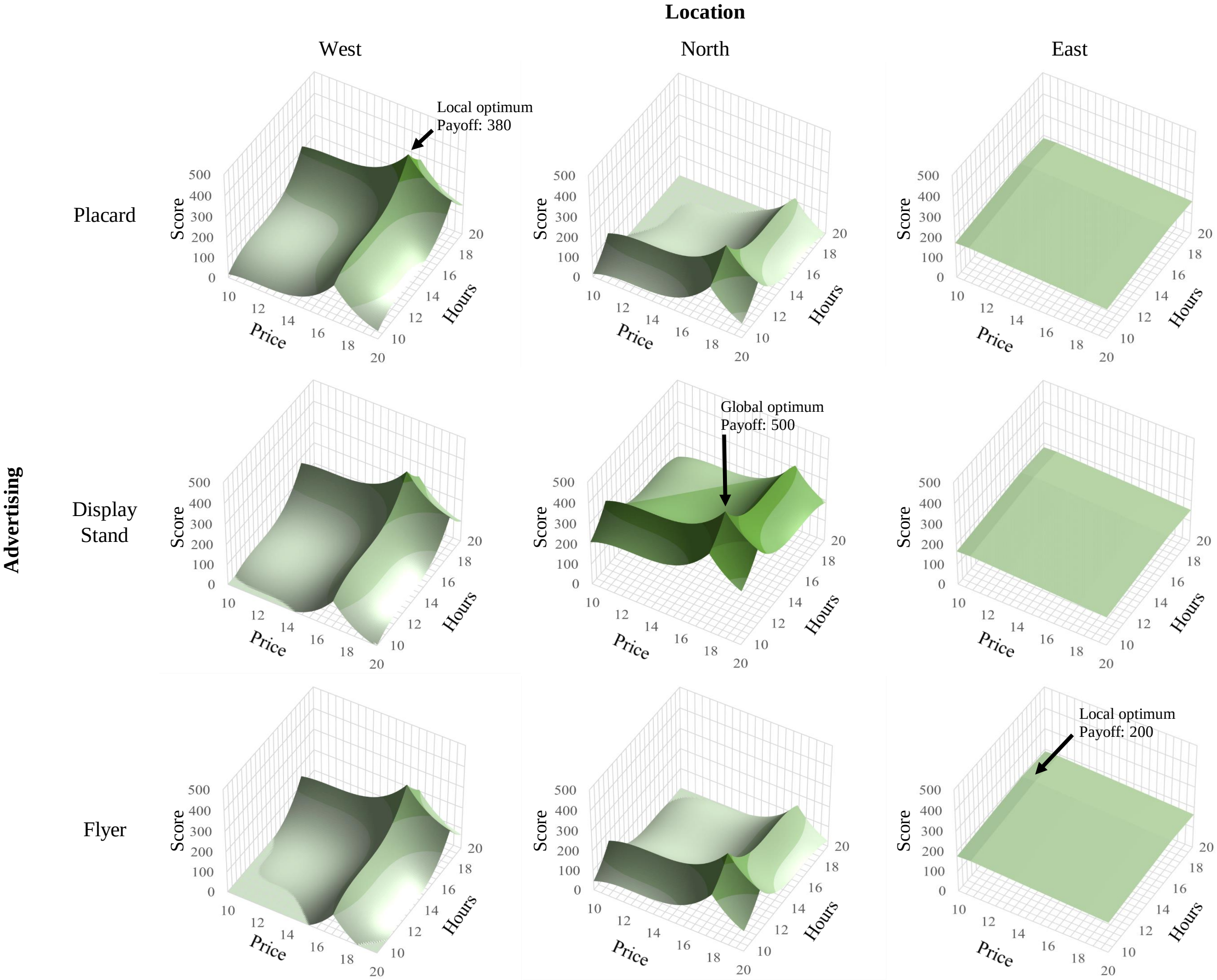
Product task:

1. Color = {Green, Yellow, Orange}
2. Lemon content = $\{10, 10.1, 10.2, ..., 19.9, 20\}$
3. Carbon dioxide content = $\{10, 10.1, 10.2, ..., 19.9, 20\}$
4. Bottle label = {Square, Triangle, Circle}

Market component:

1. Location = {West, North, East}
2. Price = $\{10, 10.1, 10.2, ..., 19.9, 20\}$
3. Opening hours = $\{10, 10.1, 10.2, ..., 19.9, 20\}$
4. Advertising = {Placard, Display Stand, Flyer}

Figure EC.2.2 Lemonade Stand Activity: Market Task (Rugged Parametrization, Version 1)

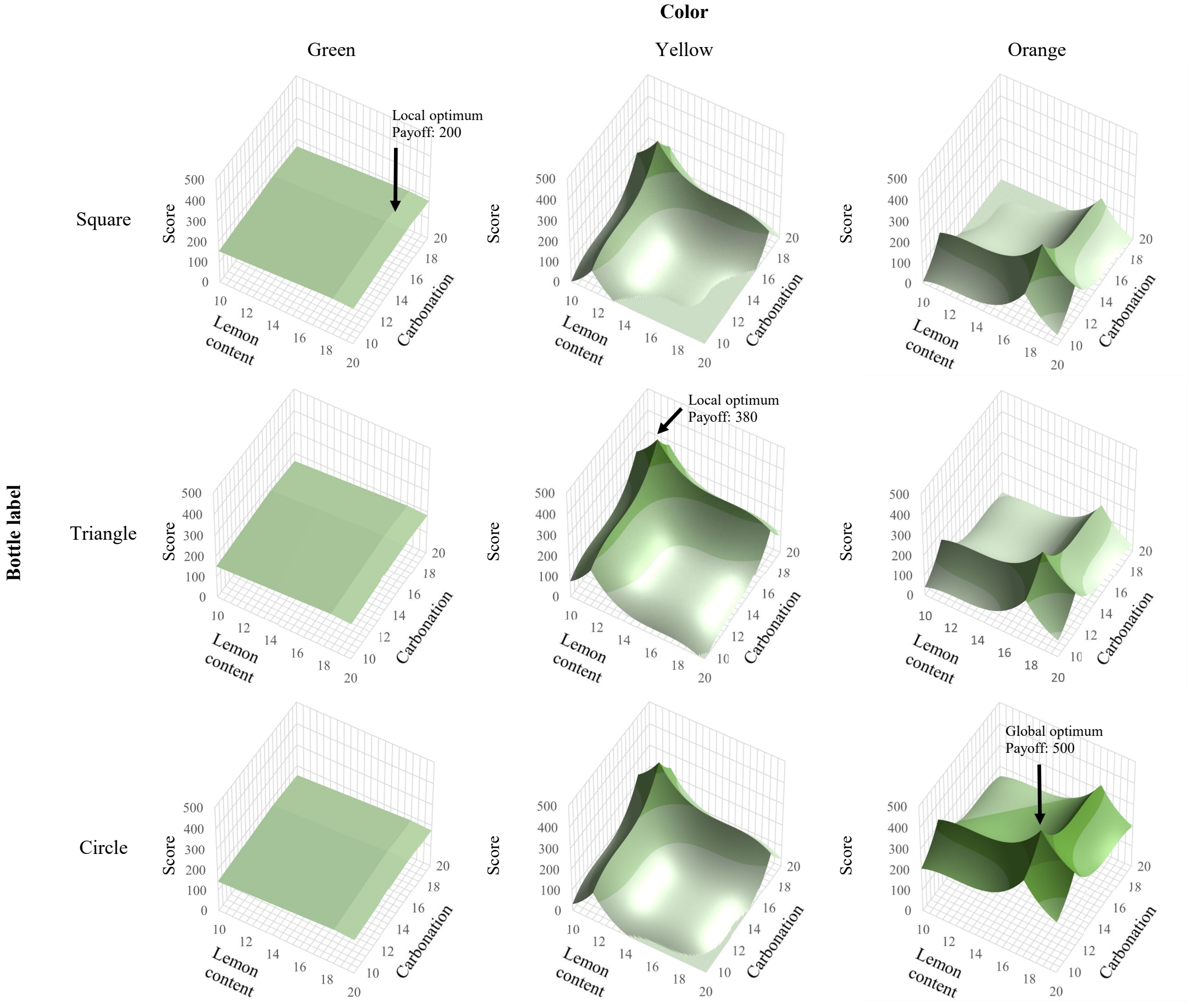


EC.2.4. Rugged Parametrization (Treatments T1-T4)

We first described the parameters used in the more rugged parametrization of the Lemonade Stand game. For each lemonade color (in the Product task) and location (in the Market task), there is a predefined, optimal selection resulting in a maximum profit. To avoid the possibility that our effects were driven by a single parameter version we used two different parameter versions for each task. Table EC.2.2 shows the optimal selections and maximum profits for each task and version.

For the market component, Figure EC.2.2 shows the three maxima, each of which corresponds to a combination of location and advertising. For the product component, Figure EC.2.3 shows the three maxima, each of which correspond to a combination of lemonade color and bottle label. As

Figure EC.2.3 Lemonade Stand Activity: Product Task (Rugged Parametrization, Version 2)



Local optimum
Payoff: 200

Local optimum
Payoff: 380

Global optimum
Payoff: 500

shown in Table EC.2.2, we set these three maxima to 200, 380, and 500 points, respectively. Note that while the optimal locations of the remaining attributes are unchanged if we move vertically in Figures EC.2.2 and EC.2.3, the locations change if we move horizontally. This corresponds to the medium complexity scenarios used in the prior rugged landscape literature (see, for example, Sommer et al. 2020, and references there). The penalties for the discrete attributes (lemonade color and bottle label for the Product component, as well as Location and Advertising for the Market component) are given in Table EC.2.3. The penalties for the lowest local maximum (at 200) for the continuous attributes (Lemon content and Carbonation for the Product component, as well as Price and Opening hours for the Market component) are linear. They were computed

Table EC.2.2 **Optimal Selection for Rugged Parametrization of Lemonade Stand Activity**

	Version 1			Version 2		
Product						
Lemonade color	Green	Yellow	Orange	Green	Yellow	Orange
Lemon content	18.5	11.6	17.6	11.5	12.4	18.4
Carbonation	16.9	18.5	12.2	13.1	17.8	11.5
Bottle label	Square	Triangle	Circle	Square	Triangle	Circle
Maximum Profit	200	380	500	380	500	200
Market						
Location	West	North	East	West	North	East
Price	17.1	17.9	10.9	12.9	19.1	12.1
Opening hours	18.5	11.8	17.3	11.5	12.7	18.2
Advertising	Placard	Display stand	Flyer	Placard	Display stand	Flyer
Maximum Profit	380	500	200	200	380	500

by multiplying each unit of absolute deviation by a constant, i.e. *absolute deviation* × 3. In order to achieve a sufficiently high level of difficulty, the penalty functions are S-shaped; that is, the gradient decreases the closer one gets to the optima. To achieve this, the penalty functions were calibrated as follows: $(\frac{absolute\ deviation}{5} - 1)^3 \times 150 + 150$. These penalty functions led to a level of difficulty that was found to be appropriate in pre-experimental pilots with 33 participants.

Table EC.2.3 **Penalties by Component and Parameter Version**

		Version 1			Version 2		
Product							
Lemonade color		Green	Yellow	Orange	Green	Yellow	Orange
Bottle label	Square	0	75	195	0	165	10
	Triangle	3	0	165	75	0	3
	Circle	10	45	0	45	195	0
Market							
Location		West	North	East	West	North	East
Advertising	Display stand	45	0	10	10	0	165
	Flyer	75	165	0	3	75	0
	Placard	0	195	3	0	45	195

EC.2.5. Smooth Parametrization (Treatments T5 and T6)

We next describe the parameters used in the smooth parametrization of the Lemonade Stand game (used in treatments T5 and T6, see Table 1 for treatment details). The parameters were chosen analogously to Table EC.2.2 and Table EC.2.3, with the difference that the optimal combination of the continuous variables in the Market task was always $Price = 18.2$ and $Hours = 11.5$, regardless of the remaining two attributes (location and advertising). Similarly, the optimal combination of the continuous variables in the Product task was always $Lemon\ content = 18.5$ and $Carbonation = 16.9$, regardless of the remaining two attributes (bottle label and color). Figures EC.2.4 and EC.2.5 show the landscapes. Note that the local optima are located in the same position in all nine combinations of discrete attributes. This is analogous to the low complexity scenario in Sommer et al. (2020).

Figure EC.2.4 Lemonade Stand Activity: Market Task (Smooth Parametrization, Version 1)

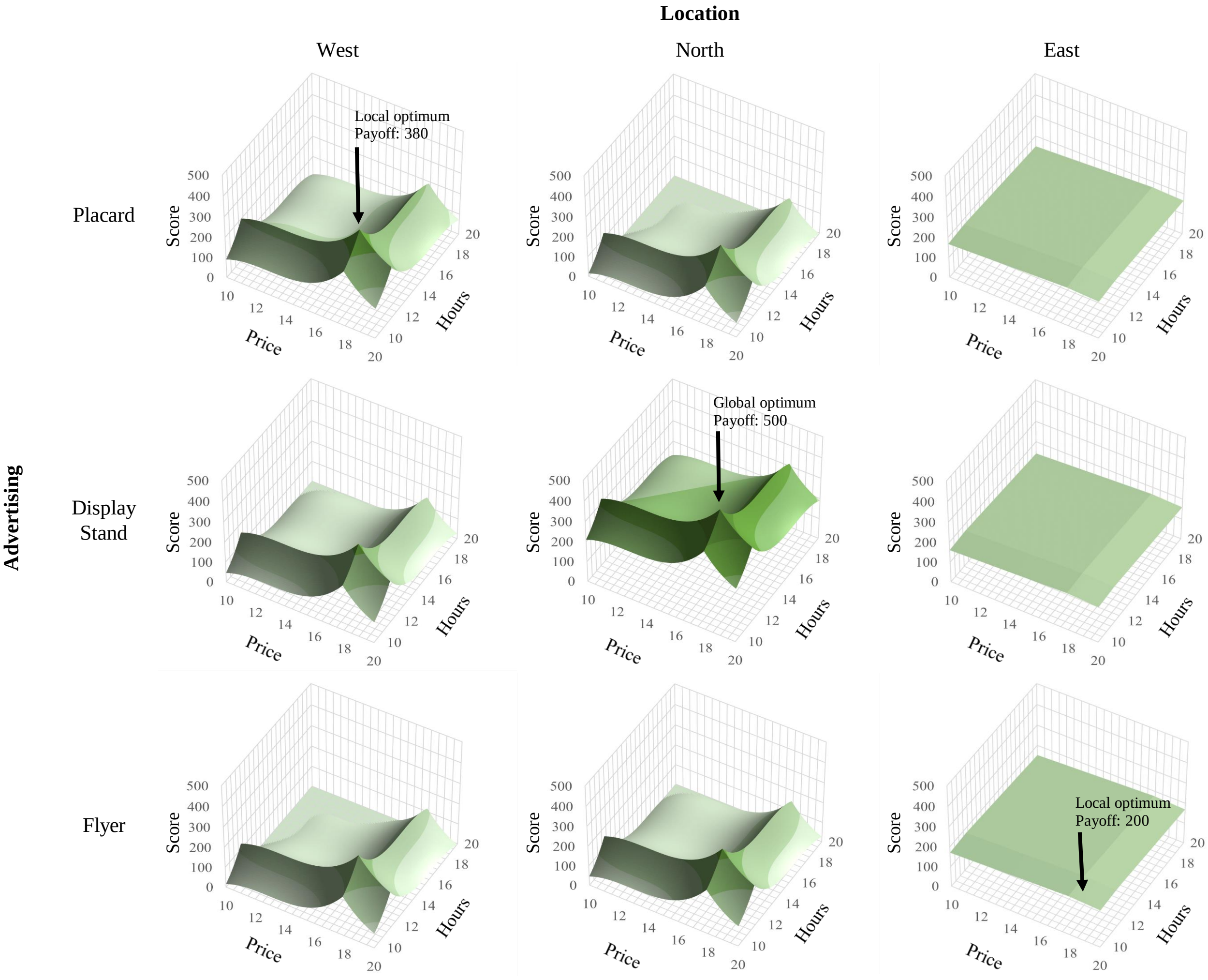
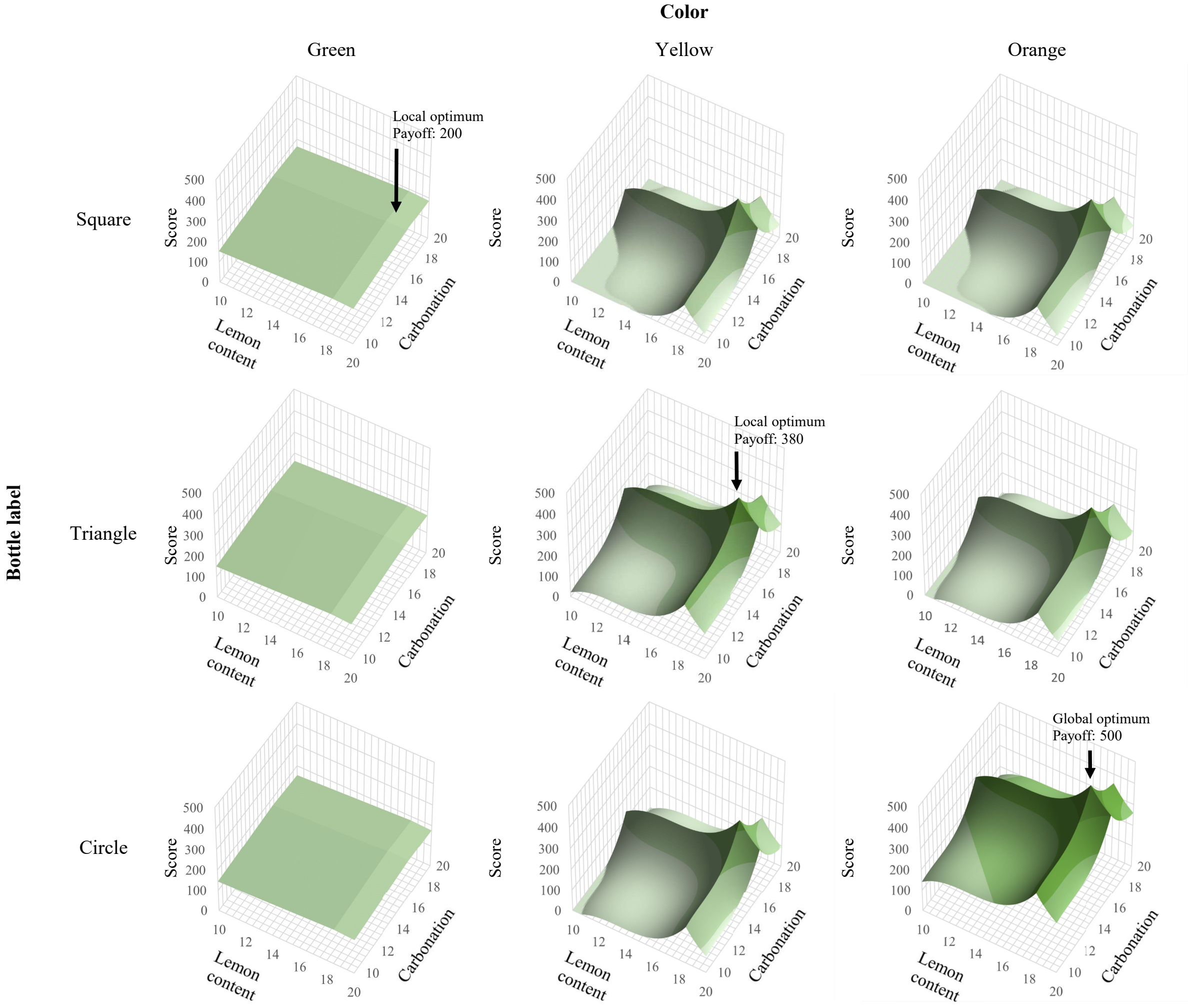


Figure EC.2.5 Lemonade Stand Activity: Product Task (Smooth Parametrization, Version 1)



EC.3. Additional Analysis

EC.3.1. Quantile Regressions of Performance on Treatments

In the main text we discussed heterogeneous treatment effects on performance. In particular, in §4.2.3 we noted that the within-treatment performance distribution looked quite similar for the right tail of the distribution, but differed between treatments for the left tail. Here we present more formal analysis of these effects. Table EC.3.4 shows the coefficients from quantile regressions of performance on treatments in the Scrabble activity. We use the same set of covariates as col. (1) in Table 3 in the main text. The analysis shows that iterative workflow improves the outcomes mainly in the low range of the performance distribution.

Table EC.3.4 Scrabble: Quantile Regressions

Quantile:	10	20	30	40	50	60	70	80	90
SEQ	Baseline	Baseline	Baseline	Baseline	Baseline	Baseline	Baseline	Baseline	Baseline
ITER EQUAL FREEZE	110.00***	65.00**	62.50**	45.00*	35.00*	35.00*	40.00*	30.00	15.00
	(26.18)	(24.22)	(22.36)	(18.89)	(17.90)	(17.93)	(19.15)	(20.56)	(27.23)
ITER FREEZE	125.00***	82.50***	67.50***	47.50**	35.00*	35.00**	25.00	15.00	-5.00
	(29.23)	(21.60)	(21.22)	(17.81)	(18.40)	(17.54)	(16.51)	(18.45)	(18.51)
ITER	95.00***	52.50**	40.00*	25.00	15.00	20.00	15.00	-0.00	-10.00
	(32.55)	(22.81)	(20.15)	(19.13)	(19.18)	(18.49)	(17.69)	(21.14)	(19.32)
Constant	35.00	87.50***	120.00***	147.50***	165.00***	195.00***	215.00***	245.00***	270.00***
	(23.97)	(19.97)	(19.02)	(16.73)	(19.04)	(20.28)	(17.66)	(18.57)	(17.11)
Observations	244	244	244	244	244	244	244	244	244

Notes: Table shows quantile regression coefficients. Dependent variable is Scrabble performance. Each column corresponds to a quantile, starting from the 10^{th} to the 90^{th} quantile. Controls are task sequence, component sequence, loss of internet connection. $^{*}p < 0.1$, $^{**}p < 0.05$, $^{***}p < 0.01$

Table EC.3.5 repeats this analysis for Scrabble with pre-formed words (§5). We use the same set of covariates as col. (1) in Table 4 in the main text.

Table EC.3.5 Scrabble with pre-formed words: Quantile regressions

Quantile:	10	20	30	40	50	60	70	80	90
SEQ	Baseline	Baseline	Baseline	Baseline	Baseline	Baseline	Baseline	Baseline	Baseline
ITER	25.00	0.00	0.00	25.00**	25.00**	0.00	0.00	25.00**	0.00
	(17.29)	(13.80)	(11.44)	(11.62)	(10.96)	(8.94)	(8.77)	(10.71)	(13.64)
Constant	25.00	50.00***	75.00***	75.00***	75.00***	100.00***	100.00***	100.00***	125.00***
	(17.14)	(11.79)	(12.13)	(8.52)	(5.17)	(10.33)	(14.20)	(12.68)	(16.26)
Observations	112	112	112	112	112	112	112	112	112

Notes: Table shows quantile regression coefficients. Dependent variable is performance. Standard errors are shown in parentheses next to the coefficients. Significance levels are denoted as $^{*}p < 0.1$, $^{**}p < 0.05$, $^{***}p < 0.01$. SEQ is considered as baseline across all quantiles.

Table EC.3.6 shows the coefficients from quantile regressions of performance on treatments, focusing on Lemonade Stand activity, rugged landscape (§6). We use the same set of covariates as in the main text. The analysis shows that iterative workflow improves the outcomes mainly in the lower range of the performance distribution.

Table EC.3.6 Lemonade Stand (Rugged Parametrization): Quantile Regressions

Quantile:	10	20	30	40	50	60	70	80	90
SEQ	Baseline	Baseline	Baseline	Baseline	Baseline	Baseline	Baseline	Baseline	Baseline
ITER EQUAL FREEZE	4.00 (4.70)	-3.00 (1.89)	-1.50 (1.26)	-0.90 (3.79)	-1.50 (26.63)	-71.45* (37.41)	-86.48** (36.83)	-59.97 (41.01)	-68.33 (38.36)
ITER FREEZE	3.10 (9.78)	-1.20 (2.04)	-1.20 (1.29)	-0.90 (3.74)	-1.50 (26.57)	-70.83 (39.14)	-73.63* (39.19)	-46.88 (41.05)	-59.42 (40.12)
ITER	10.60* (5.88)	4.50 (21.73)	71.37** (30.54)	97.53*** (25.75)	130.09*** (34.16)	61.96 (39.18)	15.89 (32.60)	22.06 (39.00)	3.98 (31.07)
Constant	187.00*** (5.04)	196.70*** (2.03)	198.20*** (1.15)	199.10*** (4.52)	200.30*** (32.48)	304.63*** (39.16)	380.00*** (31.17)	392.65*** (35.20)	465.77*** (26.03)
Observations	286	286	286	286	286	286	286	286	286

Notes: Table shows quantile regression coefficients for multiple treatments across different quantiles, with *SEQ* as baseline. Dependent variable is performance. Standard errors are shown in parentheses. Significance levels are denoted as $*p < 0.1$, $**p < 0.05$, $***p < 0.01$.

Table EC.3.7 shows the coefficients from quantile regressions of performance on treatments, focusing on Lemonade Stand activity, smooth landscape. We use the same set of covariates as in the main text. The analysis shows that iterative workflow does not affect the outcomes at any of the performance quantiles.

Table EC.3.7 Lemonade Stand (Smooth Parametrization): Quantile Regressions

Quantile:	10	20	30	40	50	60	70	80	90
SEQ	Baseline	Baseline	Baseline	Baseline	Baseline	Baseline	Baseline	Baseline	Baseline
ITER	3.90 (7.23)	7.80 (13.72)	7.50 (26.31)	48.95 (37.67)	2.62 (37.93)	-16.20 (29.51)	-3.96 (25.78)	-12.74 (36.16)	-64.78 (46.48)
Constant	198.2*** (3.54)	202.06*** (15.53)	238.3*** (27.72)	259.68*** (35.70)	322.03*** (31.71)	351.2*** (25.61)	358.21*** (25.09)	398.26*** (39.75)	479.61*** (36.92)
Observations	115	115	115	115	115	115	115	115	115

Notes: Table shows quantile regression coefficients for the treatment across different quantiles, with *SEQ* as baseline. Dependent variable is performance. Standard errors are shown in parentheses. Significance levels are denoted as $*p < 0.1$, $**p < 0.05$, $***p < 0.01$.