Pay-for-performance incentives and individual creativity: Experimental evidence

Eileen Tipoe

Abstract

Creativity is economically important, but is typically one of many goals in a business environment. Given that producing creative output involves a financial cost, a crucial question facing firms is how direct financial incentives on creativity affect cost efficiency. I use a lab experiment where participants complete a task that can vary in two dimensions (creativity and cost efficiency), and face pay-for-performance incentives specific to one dimension only. Using a novel content-based measure of creativity, I find that pay-for-performance incentives on creativity have sizeable positive direct effects without significantly affecting cost efficiency. I develop a framework that describes how individuals search the two-dimensional space for optimal solutions, and use it to estimate parameters related to search behaviour, allowing for unrestricted heterogeneity across individuals. The degree of exploration, measured in terms of the variety of output produced and the willingness to consider inferior solutions, is crucial for explaining the direct and indirect effects of goal-specific incentives.

1 Introduction

Creativity, defined as the act of coming up with unusual ideas or ways to solve a problem, plays a crucial role in modern economies (O*Net 2020). In developed countries, a significant proportion of workers have jobs that require creative problem-solving and decision-making, ranging from 20-55% of employment across OECD countries (Autor and Price 2013, Marcolin et al. 2016). Creativity is also a highly valued skill in the workplace. In Global surveys of top managers and employers reveal that fostering creativity and innovation is one of their primary concerns (The Conference Board 2014), and that creativity is one of the top 5 skills in terms of current demand and expected growth in demand over the next decade (World Economic Forum 2018).

This paper explores three crucial questions facing firms that aim to encourage creative behaviour. First, what (if any) kinds of incentives can encourage creative behaviour? Second,
how does the creative process (the development and refinement of ideas) respond to incentives? Third, given that producing creative output involves a financial cost, what (if any) are the consequences of incentivising creativity on cost efficiency?

Despite its economic importance, creativity has received limited attention from economists because it is difficult to measure creativity and cleanly identify a causal link between creativity and incentives. To address both of these issues, a common approach taken by the existing literature (including this paper) is to use controlled experiments. Regarding the question of effective incentives, there is evidence from lab experiments (Chen et al. 2012, Laske and Schroeder 2017, Bradler et al. 2019) that some financial incentive schemes such as pay-for-performance bonuses or tournaments can improve performance in real-effort creative tasks, provided that incentives are low-powered and the task is well-defined.\(^1\) However, in all studies, participants were given creativity (or specific dimensions of creativity) as the only goal,\(^2\) so it is unclear whether these results generalise to multi-goal situations, where financial incentives for creativity could affect performance on non-incentivised goals.

I make three contributions to the literature. First, I address the lack of empirical evidence on spillover effects by investigating the direct and indirect effects of goal-specific pay-for-performance incentives in a two-goal setting. To cleanly identify the effect of incentives on creative output, I use a lab experiment and a novel task that can vary in two dimensions: creativity and cost. Specifically, participants must design bridges while facing a fixed budget constraint, where each bridge component (beam or joint) has an associated cost. Participants face two goals: maximise creativity and maximise cost efficiency, defined as the amount of budget left unspent. Using a within-subject design, I randomly assign participants to one of six treatments, which vary in the incentivised goal (creativity, cost efficiency, or a flat payment regardless of performance but double the time limit), and the per-unit bonus (“low” or “high”). In contrast to commonly-used creative tasks where output is produced “for free” (aside from the necessary cognitive and physical effort),\(^3\) in my task, participants face an explicit financial cost of producing creative output.\(^4\) By adding the goal of cost efficiency, I aim to capture a key feature of real-world creative problem-solving tasks.

My second contribution is methodological. I propose a definition of creativity that allows

---

\(^1\)See Charness and Grieco (2019) for a recent review of the literature, including the effect of non-financial incentives on creativity.

\(^2\)For example, Laske and Schroeder (2017) conceptualise creativity as a multi-dimensional phenomenon consisting of quality, quantity, and originality, and examine the effects of piece-rate incentives on incentivised and non-incentivised dimensions. Kachelmeier et al. (2008) compares the effectiveness of piece-rate incentives on quantity, originality, or a weighted sum of both.

\(^3\)For example, writing a story or essay about a new invention (Charness and Grieco 2019), think of unusual uses for a common object (Bradler et al. 2019), or create rebus puzzles with no restrictions on the words or pictures used (Kachelmeier et al. 2008, Erat and Gneezy 2016).

\(^4\)While some studies add constraints to the creative task by giving participants a set of pre-defined components with which to produce output (Laske and Schroeder 2017) or components that participants must include in their answer (Charness and Grieco 2019), there is no financial cost associated with the use of each component and hence the notion of “cost efficiency” does not apply directly in these settings.
for precise measurement in the context of my experiment: distance from a pre-specified reference point. Using this definition, I measure creativity precisely with the graph edit distance metric, defined as the minimal number of operations required to transform one graph (a discrete set of dots and lines) into a given reference graph (Sanfeliu and Fu 1983). I acknowledge that this approach is only one of many possible ways to conceptualise creativity, and as with other measures of creativity, it cannot completely capture every characteristic that is associated with creativity (such as aesthetics). However, graph edit distance is still a suitable measure for my experiment task due to the discrete nature of the output, and the constraints of the bridge design task: participants must work with the components given and cannot invent new components or technologies for building bridges, so the only way to demonstrate originality is to make visually dissimilar designs. Since all conceptualisations of creativity assert that originality is a characteristic feature of creative output,\(^5\) graph edit distance can be interpreted as a measure of creativity in my specific context.

Using this measure of creativity, I find that pay-for-performance incentives on creativity have sizeable positive direct effects without associated changes in cost efficiency: on average, an individual’s most creative design in the treatment condition is significantly more creative than their most creative design in the control condition, but does not substantially differ in cost. One explanation for this significant direct effect is that the variety of designs (measured by the standard deviation of the edit distance) is significantly correlated with performance in the creativity goal, and participants tried a wider variety of designs in the treatment condition.

The multiple treatment conditions allows me to compare the cost effectiveness of different pay-for-performance schemes, which is important from a business perspective. Comparing the cost (to the researcher) of inducing creativity across all treatments and the control condition, high per-unit incentives on creativity has the lowest average cost (£ per person) and highest amount of creativity per pound spent, making it the most cost-effective scheme.

My third contribution is a framework for thinking about how individuals develop and refine ideas, referred to as “the creative process”. In the context of my experiment, participants make designs with the aim of maximising utility \(U\), defined as a weighted sum of the design’s creativity \((c)\) and cost efficiency \((e)\):

\[
U = \theta c + (1 - \theta) e
\]

where the relative weight on creativity, \(\theta \in [0, 1]\), varies across individuals. Given the wide variety of bridge components and possible designs, the solution space is characterised by many local maxima rather than a unique global maximum. However, participants do

\(^5\)For example, Laske and Schroeder (2017) identify originality as a key dimension of creativity, and Gross (2020) considers an idea to be creative if it is “novel and appropriate to the goal at hand”. See Section 2.1 for a detailed discussion on definitions of creativity.
not know the set of possible solutions \textit{ex ante}, so must search this solution space by trying different designs in the allotted time. I outline an algorithm that describes a plausible way for participants to search this space, which depends on two key parameters: the relative weight on creativity, and the willingness to consider initially inferior solutions (designs that give less utility than the current design but may lead to a better solution). This algorithm constitutes a discrete-time inhomogeneous Markov chain with one-step transition probabilities, and is based on the concept of simulated annealing (Kirkpatrick et al. 1983, Cerny 1985). While simulated annealing has been used for other purposes in the economic literature, this is the first paper to use simulated annealing as a model of human behaviour instead of a method of maximisation.

An important feature of my experimental setup is that I can observe the entire sequence of designs submitted by each participant. I can therefore estimate the algorithm’s parameters at the individual level, allowing for unrestricted heterogeneity across individuals. I find that the average willingness to consider inferior solutions increases under high-powered incentives for creativity, but decreases under high-powered incentives for cost efficiency. The willingness to consider inferior solutions is also significantly positively correlated with performance in the creativity goal whereas the relative weight on creativity is not, suggesting that how participants tried to maximise utility is more important for outcomes than what they tried to maximise. Taken together, my results indicate that the creative process is important for understanding the relationship between incentives and outcomes, and that exploration (trying a variety of solutions and being willing to consider inferior solutions) plays a crucial role in creative task performance.

The remainder of this paper is organized as follows. Section 2 provides an overview of the related literature. Section 3 describes the experimental design. Section 4 outlines the methodology for measuring of creativity and the framework for characterising the creative process. Section 5 summarises the experimental results on the direct and indirect effects of incentives, and explores the relationship between the creative process and task performance. Section 6 discusses the main findings and Section 7 concludes.

2 Related literature

This paper draws together the literature on financial incentives, the creative process, and creative outcomes by exploring the interplay between these three aspects.

2.1 Defining and measuring creativity

The first attempt to formally define and measure creativity comes from the psychologist Torrance (1969) who defines creativity as “the capacity to detect gaps, propose various solutions
to solve problems, produce novel ideas, re-combine them, and intuit a novel relationship between ideas.” The Torrance test of creative thinking (Torrance 1962, 1974, 1989), which consists of a series of verbal and non-verbal activities, is still the most commonly used way to measure creativity in the psychology literature (Almeida et al. 2008). Since Torrance’s seminal work on creativity, psychologists have sought to define creativity more precisely, though all definitions assert that novelty and usefulness are key characteristics of creative output. For example, Stein (1974) and Woodman et al. (1993) define creativity as “the production of novel and useful ideas in any domain”. The economic literature on creativity also uses these definitions.\(^6\)

Following this approach, I define creativity as the distance from a pre-specified reference point, which captures the notion of novelty in the context of my bridge design task. Since participants must make designs with the components given and do not have the scope to invent new components or technologies for building bridges, visual dissimilarity is the only way to judge the novelty of a design. I present evidence in Section 4.1 that this definition of creativity, as measured by the graph edit distance metric, is a plausible method of evaluating visual dissimilarity.

An important distinction to make at this point is how creativity (as defined here) differs conceptually from innovation, as these terms are sometimes used interchangeably in layman speech. While both concepts involve experimentation, the distinguishing feature of innovation is the production and implementation of new knowledge (Arrow 1969, Weitzman 1979). Creativity refers to an output’s uniqueness or deviation from existing outputs but does not necessarily require new knowledge, and thus is considered as a precursor of innovation.

In the experimental literature, creativity is commonly measured on a categorical scale by a group of external judges. With the exception of some specific tasks (such as the verbal task of Charness and Grieco (2019)) judges are usually not given further instructions on how to rate the creativity of an output. More recently, there have been some developments in the use of technology to aid evaluations of creativity, such as in the field study of Gross (2020), which uses image comparison algorithms to assess the similarity of logo designs. In Section 4.1, I build on this recent literature by presenting an alternative method of measuring creativity (graph edit distance) that is particularly suitable for my experiment task.

---

\(^6\)For example, Charness and Grieco (2019) use the definition of Stein (1974) and Woodman et al. (1993), Laske and Schroeder (2017) define creativity as a “multidimensional phenomenon” that consists of quality, originality, and quantity, and Gross (2020) defines creativity as “the act of producing ideas that are novel and appropriate to the goal at hand”.

---

5
2.2 Financial incentives and creativity: Theory and experimental evidence

The theoretical consensus in the economics and psychology literature is that financial incentives are detrimental to creativity, for several reasons. Firstly, monetary incentives could potentially crowd out any intrinsic motivation that individuals have to produce creative output, and thus be counterproductive. The experimental studies of Deci and Ryan (1985), Amabile (1996), and Gneezy and Rustichini (2000) supports this hypothesis. Another argument is that monetary incentives increase the consequences of poor performance in the task, causing individuals to “choke” under pressure (Ariely 2009).

However, other studies find that financial incentives can be effective in certain contexts. For example, Charness and Grieco (2019) provide evidence that financial incentives in the form of tournament competition can increase creativity if the creative task has a specific and well-defined ex ante goal (“closed” tasks), though not if the task lacks an underlying ex ante goal (“open” tasks). The task I use in my experiment is “closed” (find a solution to a specific problem), and I build on their findings by considering the effect of other financial incentives (pay-for-performance) when there is another goal besides creativity.

In multidimensional settings, economic theory predicts that if the dimensions are substitutes in effort costs, incentivising specific dimension(s) will have negative spillover effects on the non-incentivised dimension(s) (Holmstrom and Milgrom 1991). Both Kachelmeier et al. (2008) and Laske and Schroeder (2017) conduct lab experiments where only one dimension of creativity (quality or quantity) is incentivised, and find evidence of a quality-quantity tradeoff. While creativity and cost efficiency are plausibly substitutes (making a more-creative design might require more materials and hence a higher cost), I do not find evidence of negative spillover effects in my experiment, which may be because of complementarities in characteristics of search behaviour (for example, a greater willingness to experiment is positively correlated with higher creativity as well as lower cost).

2.3 Incentives and the creative process

Another branch of the literature explores the relationship between financial incentives and the way that individuals search for and refine their solutions (the “creative process”). The psychology literature argues that incentives can be detrimental to creativity because they cause individuals to narrow their focus of attention, resulting in a smaller solution set being considered (Easterbrook 1959). Financial incentives can also distract individuals from the task (such as considering what would happen if they did or did not get a reward), especially if the incentives are large (Ariely 2009).

There is some experimental evidence that financial incentives can discourage the exploration of new approaches. In the field experiment of Englemaier et al. (2018), financial incentives
decreased the likelihood of exploration, but only for teams who had lower intrinsic motivation (were mandated to perform the task). In a field study of a logo design competition, Gross (2020) finds that moderate levels of competition are positively associated with exploration, but high levels of competition discourage exploration (individuals stop competing). Ederer and Manso (2013) find that the structure of financial incentive schemes can affect search behaviour: in a multi-period profit-maximisation simulation, individuals under a payment scheme that tolerates early failure and rewards long-term success are more likely to explore the solution space, resulting in better performance compared to fixed-wage or standard pay-for-performance incentive schemes. My experiment task is a simulation similar to that of Ederer and Manso (2013), but differs crucially in that individuals do not have a one-dimensional goal. Unlike their simulation, my task has multiple local maxima rather than a single correct solution (global maximum), so allows for a wider range of creative expression.

3 The experiment

My experiment used bridge design as the creative task. Participants used a simulation program to design bridges according to specified criteria. I use a $2 \times 3$ within-subject design (Figure 1), with each participant completing a control task and one of six randomly assigned treatments, which vary according to the type of goal incentivised (creativity, cost efficiency, flat payment regardless of performance), and the size of the per-unit bonus (“high” or “low”, measured in units of creativity, dollars, and time respectively).

In line with my definition of creativity, participants were told that creativity would be measured in terms of visual dissimilarity from a specified reference bridge. Participants were not told exactly how visual dissimilarity would be evaluated, but were given one example of a less-creative (“quite similar”) output and one example of a more-creative (“less similar”) output (see Appendix A for the full instruction sheets). In all treatments, participants were told that they could make as many designs as they wished in the allotted time. Participants who were paid according to creativity or cost efficiency were also told that their bonus would only depend on the most creative design or cheapest design, respectively.
3.1 Task description

Participants used a bridge-building simulation with the aim of designing bridges that were as creative as possible, while also staying within a budget of $400,000. The bridges also had to be structurally stable, meaning that a standard-size (480kN) truck could successfully drive across the bridge. Participants could test the structural stability of their current design at any time (by pressing a button in the program), and their designs were saved automatically after each test. Figure 21 (Appendix C) shows the simulation interface and the information that participants know for certain at any point in time: the cost and structural stability of their design.

A bridge design consists of two basic elements: joints (circles) and beams (lines). Participants could use as many of each element as they wished (subject to the budget constraint), as well as change the material (3 possible types), cross-section (hollow or solid) and thickness of the beams (33 possible options). The task therefore has a vast range of possible solutions rather than a single correct or globally optimal solution.

To familiarise themselves with the program, participants worked through a short tutorial and were given 15 minutes to work on a practice template before starting the experiment. To prevent learning across treatment and control conditions, participants were given a different template for each condition (Figure 2), assigned in a random order. Each template requires a different type of design to make a structurally stable bridge, so participants could not simply re-use their designs from the previous task.
Figure 2. Templates given to participants

The top panel shows the practice bridge template. The bottom left panel shows the truss bridge template (“Template 1”), which requires the bridge to have a flat base. The bottom right panel shows the arch bridge template (“Template 2”), which requires the bridge to have supports in the base. Unfilled circles indicate where the nodes for the base should be positioned in order to build a stable bridge with the least number of nodes possible.

3.2 Treatments

There were six treatment conditions and one control condition. Each participant completed the control condition and one treatment condition (presented in a random order).

Control. Participants were paid a flat payment of £2.50 and given a time limit of 20 minutes to work on the task.

1. Flat payment, low rate (Extra-time-low). Participants were paid a flat payment of £2.50 and given a time limit of 40 minutes to work on the task.

2. Flat payment, high rate (Extra-time-high). Participants were paid a flat payment of £5.00 and given a time limit of 40 minutes to work on the task.

3. Creativity, low rate (Creativity-low). Participants were paid between £1.50 and £3.50 (£0.02 per unit of creativity), and given a time limit of 20 minutes to work on the task.

4. Creativity, high rate (Creativity-high). Participants were paid between £0.50 and £4.50 (£0.04 per unit of creativity), depending on the creativity level of their most creative output, and given a time limit of 20 minutes to work on the task.

5. Cost efficiency, low rate (Efficiency-low). Participants were paid between £1.50 and £3.50, proportional to the remaining budget from their cheapest bridge (£0.02 per $1000 remaining), and given a time limit of 20 minutes to work on the task.
6. Cost efficiency, high rate (Efficiency-high). Participants were paid between £0.50 and £4.50, proportional to the remaining budget from their cheapest bridge (£0.04 per $1000 remaining), and given a time limit of 20 minutes to work on the task.

For the creativity and cost efficiency treatments (3-6), participants were told that the bonus only depended on their own performance (to eliminate the notion of peer competition) and had been “calibrated according to the parameters of the bridge template” so that the average total payment for this task would be £2.50 (the same payment as the control condition). The per-unit payments used are within the range of the piece rates used in previous studies.

Participants were also told that only designs that satisfied both the cost and structural stability criteria would be considered for the bonus; if none of their designs met both criteria they would be paid £0.50 (for the high rate treatments) or £1.50 (for the low rate treatments). Thus, the payment in the “low” treatments (3 and 5) ranges from 60-140% of the payment in the control condition, and payment in the “high” treatments (4 and 6) ranges 20-180% of the payment in the control condition. The relative size of bonus to baseline payment is within the range used by previous studies.

3.3 Exit survey

After completing both tasks, participants were asked to provide demographic information, and also answered questions on their:

1. Willingness to take risks, measured using a 0-10 Likert scale (Dohmen et al. 2011),
2. Approach to problem-solving, ranging from systematic to complete trial and error (Nielsen et al. 2008),
3. Big-5 personality traits, using the 15-item questionnaire of Lang et al. (2011),
4. Tendency to persist on a task until it is complete (“grit”), using the 8-item questionnaire of Duckworth (2007),

---

7 The per-unit payments (£0.02 and £0.04) were determined by dividing the theoretical maximum amount of the variable of interest (100 for creativity and $100,000 for cost efficiency, estimated via an initial pilot study) equally across the specified range. For example, the per-unit payment for creativity (low) would be $(3.00 - 1.00)/100 = £0.02 per unit of creativity.

8 For example, Laske and Schroeder (2017) give a base payment of 2.50 Euros (£2.24) and a payment of 0.1 Euros (£0.09) per unit of creativity (where units ranged from 1 to 10); Bradler et al. (2019) give a base payment of 3 Euros (£2.69) plus a piece-rate payment of 0.05 Euros (£0.045) per idea.

9 For example, Eckhart et al. (2012) give 10 Euros in the control condition and 5-25 Euros in the treatment condition (a relative range of 50-250%); Charness and Grieco (2019) give $14 in the control condition and $8-20 in the treatment condition (a relative range of 57-143%).

10 Participants were asked “How willing at you to take risks, in general?” This measure was chosen instead of eliciting risk attitudes with a paid lottery because it is has been shown to be a behaviourally valid measure of risk attitudes, both as a reliable predictor of actual risky behaviour across multiple domains, and actual choices in paid lotteries, and is comparatively easier for participants to understand and cheap to administer (Dohmen et al. 2011).
5. Enjoyment of the task on a 1-5 Likert scale (a proxy for intrinsic motivation),
6. Exposure to relevant knowledge via formal learning (either in the last four years of secondary school or in university), and
7. Exposure to similar tasks such as simulation games that involve design.
These characteristics were chosen because previous studies have identified them to be strong correlates of performance in other creative tasks, or important dimensions of heterogeneity in task performance.\textsuperscript{11} The full questionnaire is shown in Appendix A.

3.4 Conducting the experiment

The experiment was conducted on Prolific, a reputable site with a large participant pool (over 100,000 active subjects), used primarily by researchers for surveys and experiments.\textsuperscript{12} There were a total of 550 participants across 10 sessions: 87 in creativity-high, 91 in creativity-low, 96 in efficiency-high, 93 in efficiency-low, 95 in extra-time-low, and 88 in extra-time-high. Participants were drawn from Prolific’s demographically diverse participant pool, with eligible participants being English speakers above the age of 18 (inclusive).\textsuperscript{13} Participants were paid a show-up fee of £2.50, plus any payments they received from the treatment condition. Before each task, participants were given written instructions, as well as some questions about the task (such as which goal was incentivised and what was the time limit) that they had to answer correctly before they could begin working on the task. About the instructions before they could start the task. The average completion time was 70 minutes, with average total earnings of £5.80. Since the experiment was conducted over the course of a week, all bonuses were paid after the final participant completed the experiment, so even if participants were able to communicate with each other via message boards, they could not pass on any feedback about which designs were paid more highly.

\textsuperscript{11}For example, Charness and Grieco (2019) find that exposure to similar tasks is negatively correlated with performance on tasks with well-defined goals (“closed” tasks), and problem-solving approaches that are more experimental (trial and error) than systematic are positively correlated with performance on “closed” and “open” tasks.
\textsuperscript{12}Studies have compared Prolific to other similar platforms such as MTurk as well as university subject pools, and found that Prolific could replicate existing results from in-person experiments and also delivered higher or comparable data quality to all comparison methods (Peer et al. 2017, Palan and Schitter 2018).
\textsuperscript{13}See \url{https://www.prolific.co/demographics/} for the demographic breakdown of Prolific participants. All participants are aged 18 or above, and the majority live in the UK or the US (40% and 32% of all subjects, respectively).
4 Methodology: Measuring creativity and modelling the creative process

4.1 Measuring creativity

As discussed in Section 2.1, creativity has been conceptualised in various ways, but all conceptualisations agree that a characteristic feature of creative output is its originality, and the degree of creativity increases in originality.\textsuperscript{14} In mathematical terms, the notion of “more original” can be thought of as the distance from a pre-specified reference point, a plausible candidate being the common or “standard” solution. I compare the bridge designs to the pre-specified bridge design for that template (Figure 3), which are provided by the simulation program.\textsuperscript{15}

\textbf{Figure 3.} Pre-specified bridge designs.

![Pre-specified bridge designs](image)

The left panel shows the pre-specified bridge for Template 1; the right panel shows the pre-specified bridge for Template 2. Both of these bridges were provided by the simulation program and are considered as the reference points for this experiment.

To evaluate the creativity of the bridge designs, I use a measure called the graph edit distance (henceforth referred to as “edit distance”), defined as the minimal number of operations required to transform one graph (a discrete set of dots (nodes) and lines (edges)) into another graph. This measure satisfies all properties of a metric (Sanfeliu and Fu 1983), and is suitable for the experiment task because the bridge designs can be summarised as a graph, where joints are the nodes and beams are the edges.

There are three possible types of operations: add or remove a beam, or move an existing beam (illustrated in Figure 4). Since each beam in a bridge must be connected to at least one node (and vice versa), other graph transformation operations such as moving nodes are

\textsuperscript{14}Commonly-cited definitions of creativity in both the economics and psychology literature include “the production of novel and useful ideas in any domain” (Stein 1974, Woodman et al. 1993), and “the act of producing ideas that are novel and appropriate to the goal at hand” (Amabile 1996, Sternberg 2008).

\textsuperscript{15}Note that the pre-specified design is not the most cost-efficient design for that template, so participants incentivised for cost efficiency have scope to improve on that design. Also, creativity is defined in relative terms, according to existing outputs, so even though the pre-specified design could have been considered creative at the time of its invention (relative to the “standard” or “common” designs at that time), future designs must be compared to the (now) standard design.
not necessary in this context. All other possible changes to the design, such as modifications to the material, cross-section, or thickness of beams, are not counted in the edit distance metric, because these changes affect the bridge’s structural stability but not its appearance. Creativity in these aspects, such as inventing new bridge-building materials, is beyond the scope of this paper.

**Figure 4.** Possible graph transformation operations

(a): move an existing beam; (b): adding or removing a beam; (c): add or remove a node.

Figure illustrates how the edit distance was calculated. The total edit distance is 8, consisting of 2 beams added (shown in green) and 6 beams moved (shown in red). In contrast with other commonly-used measures of creativity, which require manual evaluation, edit distance can be calculated automatically using adjacency matrices (see Appendix B for full details of the methodology). The use of technology to aid evaluations of creativity follows a similar principle to Gross (2020), who uses image comparison methods to assess logo designs.

**Figure 5.** Example of the edit distance calculation

The left panel shows the reference bridge. The right panel shows a candidate bridge for comparison. Red lines indicate beams that moved position (6 total); green lines indicate beams that were added (2 total). No nodes were added or removed. The total edit design is the sum of all changes = 6 + 2 = 8.
Advantages and limitations of edit distance as a measure of creativity

The edit distance has a number of advantages compared to the 1-10 scale commonly used in other creativity studies. For a given standard bridge and a given candidate bridge, the edit distance is deterministic (gives a single number), reducing the possibility of measurement error. The edit distance can also take on a wider range of values, enabling a complete ranking over designs even in cases where designs are quite similar (and may have been given the same numerical rating on a 1-10 scale). The edit distance also has a straightforward interpretation because it varies linearly with the number of design changes, so a design that scores 10 has twice the number of changes as a design that scores 5.

Figure 6 shows examples of actual output produced by participants and the associated cost of production. This figure suggests that edit distance is a sensible measure of visual dissimilarity, which, as previously mentioned, is the only dimension which we can use to evaluate a design’s originality, given that participants cannot create new bridge-building materials or technologies. Comparing the cost of producing designs with higher edit distance, one point to note is that there is no mechanical relationship between edit distance and cost: while edit distance is positively (but not perfectly) correlated with the complexity of a design, it is possible to increase a design’s edit distance without necessarily increasing cost.\footnote{Recall that changing the position of existing nodes and beams also counts as a transformation operation, so edit distance can increase without having to add new nodes and beams.}

One limitation of edit distance is that it cannot completely capture every characteristic that is associated with the creativity of a bridge design. Some of these characteristics, such as the use of innovative materials or combinations of materials, and invention of new materials or technologies used to build the bridge, are beyond the scope of the experimental task. Given the constraints of my task, visual dissimilarity is the only way that designs can vary in originality, so creativity necessarily takes a narrower definition than the typical conceptualisations used in previous studies.

Edit distance also does not capture other characteristics that contribute to a design’s value, such as aesthetics. However, incorporating this notion into a precise measure of creativity is challenging, because even if the concept of “aesthetically pleasing” could be precisely measured (for example, by the degree of symmetry), there is no ex ante guidance on how aesthetics should be weighted relative to a design’s edit distance. If two designs have the same edit distance, it would be reasonable for the “more aesthetically pleasing” design to have a higher creative score than the “less aesthetically pleasing” design, but how much higher should the score be? In order to avoid this debate and dispel the notion that visual difference (as measured in my experiment) involves aesthetic judgements, in the instructions I clarify that “‘different’ is taken in the literal sense, so bridges that are asymmetrical or irregularly shaped in any way are also considered different”. 

16
This figure shows that the edit distance is a reasonable measure of visual dissimilarity (which, given the constraints of the design software, is the only dimension through which designs can vary in creativity). Edit distance is defined as the minimal number of operations to transform one output (bridge designed by a participant) into the pre-specified reference bridge. In both panels, the second column shows examples of output produced by different experiment participants, ranked in order of edit distance (top = lowest edit distance, bottom = highest edit distance). The relevant reference bridge is shown in the first column. Participants were told to make bridges that “look as different from the [reference bridge] as possible”, but were not told exactly how visual dissimilarity would be evaluated.

Another limitation is that this measure does not include changes that affect the bridge’s appearance without altering its graph, for example dragging a node (and its connected beams) to a different position. These changes could be considered as relatively minor alterations ("tweaks") that involve a lesser degree of creative thinking compared to changing the structure of the bridge itself.\textsuperscript{17} Appendix B summarises the challenges with measuring design tweaks and incorporating them into the edit distance measure.

I present some evidence that edit distance adequately captures visual dissimilarity in spite of

\textsuperscript{17} Appendix B outlines how to extend the edit distance metric to incorporate design tweaks.
these limitations. In the exit survey (after completing all design tasks), I asked participants to rank a pre-defined set of bridges (drawn by myself) that vary according to edit distance, based on their opinion of visual dissimilarity (which is analogous to creativity in the context of my experiment). Figure 7 shows that the mean rank of each bridge (averaged across all participants) exactly corresponds to the actual rank based on edit distance (the median rank also follows the same order). The majority of participants (61%) got the exact ranking correct.

**Figure 7.** Comparison of edit distance and visual dissimilarity, using participants’ evaluations

![Bridge comparison diagram](image)

This figure provides evidence that edit distance is a plausible measure of visual dissimilarity. In the context of this task, visual dissimilarity is analogous to creativity because it is the only dimension that designs can vary in originality. Participants were given a picture of one of the pre-specified “standard” bridges (Template 1) and asked to rank the bridges in the left panel “according to how different you think they are from the standard bridge”, where 1 = most different, and 4 = least different. Bridges were given to participants in a random order. The right panel compares the average rank assigned to each bridge (x-axis) with the actual rank of that bridge, measured according to edit distance (y-axis). Error bars indicate 95% confidence intervals. All differences in means are significant at the 5% level, suggesting a clear consensus over the bridge rankings.

### 4.2 Modelling the creative process

My experimental setup allows me to observe the entire sequence of designs, enabling me to investigate the effects on pay-for-performance incentives on the way that individuals search for possible solutions.

---

18 The ranking according to edit distance (top to bottom) is 1, 2, 3, 4. If tweaks were added to the measure of creativity (see Appendix B for a definition), then the ranking would be 2, 1, 3, 4 (if tweaks are defined as nodes that moved without changing the original relationship between beams and nodes) and 2, 1, 4, 3 (if tweaks are defined as beams that changed length without changing the original relationship between beams and nodes). Though the rankings are similar, the fact that the mean and median exactly correspond to the edit distance ranking supports my use of edit distance alone as a measure of creativity.
Figures 8 and 9 show two examples of the approaches taken by individuals in my experiment. In Figure 8, the participant initially explored a variety of designs before settling on a particular design. In contrast, the participant in Figure 9 worked on many designs that were similar to the pre-specified reference bridge (low edit distance), then worked on a number of designs that were more different, and ultimately returned to their initial design.

**Figure 8.** Example #1: Sequence of designs submitted by a participant

The top panel shows the actual designs submitted by a participant, with the order of submission (iteration number) shown underneath each design. The bottom panel shows the creativity and cost efficiency ($000s, rescaled) of each design. Creativity is measured using edit distance, defined as the minimal number of operations required to transform one bridge design into the pre-specified reference bridge design. Cost efficiency ($000s) is measured as the amount of budget remaining, shifted vertically upwards so that all values are positive.
Tables 4 and 5 (Appendix C) present some reduced-form evidence that participants’ search behaviour is path dependent, but with a limited memory: both the change in creativity (edit distance) and cost are significantly correlated with characteristics of the previous design (structural stability, remaining budget, level of creativity, design number), but subsequent lags of the independent variables (2 or 3 previous periods) are not statistically significant. The negative coefficients on design number suggest that as individuals produce more designs, the designs become more similar to the previous design in terms of creativity and cost, suggesting that individuals are “narrowing” their search for possible solutions.

To check for heterogeneity in coefficients across treatments, I conduct a Wald test with the null hypothesis being equality of coefficients across all treatments. The null hypothesis is rejected at conventional significance levels for most variables in the creativity regressions but can only be rejected for design number ($p = 0.0001$) and cost efficiency of the previous
design ($p = 0.063$) in the cost efficiency regressions. Therefore there is some evidence of heterogeneity in search behaviour across treatments.

### 4.2.1 Utility-maximisation framework

I outline a framework that captures the features of short-memory path-dependent search described in the previous subsection. Each individual seeks to maximise utility, defined as a weighted sum of creativity ($c$) and cost efficiency ($e$):

$$U = \theta c + (1 - \theta)e$$

$\theta$ represents the weight that the individual assigns to creativity, and can vary across individuals and tasks. For computational tractability, $\theta$ is normalised to be in the interval $[0, 1]$ for all individuals.

The individual has a finite number of periods to search ($t = 1, ..., T$), where each period corresponds to a submitted design. When individuals build their design, they observe the realised values of creativity and cost efficiency associated with that design ($c_t, e_t$).

I acknowledge that in my experiment setup, individuals observe the realised cost efficiency of their designs but not the realised creativity, and so may be maximising according to a different notion of creativity instead of edit distance. However, this limitation is common for creative tasks that do not have a unique solution: individuals rarely receive feedback on how creative their output is until the experiment is over.\(^{19}\) I assume that individuals observe a noisy measure of creativity with mean-zero classical measurement error: $\tilde{c}_t = c_t + \varepsilon_t$, so they observe the true creativity $c_t$ in expectation. This assumption is plausible because the majority of individuals’ subjective opinions about visual dissimilarity (the measure of creativity in my context) were consistent with the rankings given by the edit distance metric (Figure 7, Section 4.1).

I take the first design (at $t = 1$) as given, because every individual should submit at least one design. Suppose that in each subsequent period, the individual uses the following algorithm to search the solution space:

1. Choose a design at random, where each possible design in the solution space has an equal likelihood of being chosen.
2. Evaluate the utility function based on the realised value of $c$ and $e$: $U_t = \theta c_t + (1 - \theta)e_t$.\(^\text{19}\)
3a. If utility weakly increases ($U_t \geq U_{t-1}$), continue working on the design in the next period. Then return to Step 2.
3b. If utility decreases ($U_t < U_{t-1}$), continue working on the design in the next period with positive probability $t^{-\frac{(U_{t-1} - U_t)}{d}}$, where $d$ is a strictly positive constant. Otherwise, return to Step 1.

---

\(^{19}\)One exception is the logo design competition of Gross (2020), where individuals observe a noisy measure of how creative their design is (the logo sponsor’s rating on a 1-5 scale).
Figure 10 provides a flowchart representation of this algorithm. Ceteris paribus, the probability that an individual continues to work on an “inferior” design decreases with time \( (t) \) and the difference in utility compared to the previous period, and increases with the strictly positive constant \( d \). I define \( d \) as the willingness to consider inferior solutions. Since the solution space is characterised by multiple local maxima, if individuals are willing to work on designs that are initially inferior to their current maxima, they may discover other maxima that give higher utility.

This framework has two observable implications. First, this framework predicts that individuals will search the solution space more widely in initial periods and gradually narrow down their search over time. Second, since the probability of continuing to work on an inferior design in the next period is strictly positive, individuals will occasionally try designs that differ radically from their current design.

**Figure 10.** Flowchart of utility-maximisation search algorithm

This flowchart summarises the proposed utility-maximisation algorithm. Boxes indicate action points, arrows indicate the transitions from one activity to another, and text over the arrows indicate the conditions underlying each transition. Individuals seek to maximise their utility function, \( U \), which is a weighted average of creativity and cost efficiency. They search in discrete time and submit one design per period, so \( t \) represents both the time period and the design number. \( d \) is a positive constant that represents an individual’s willingness to consider inferior solutions (continue working on designs that give lower utility than the previous design).

An important point to note about this framework is that it is based on the simulated annealing algorithm (Kirkpatrick et al. 1983, Cerny 1985) but uses this algorithm as a model of human behaviour rather than its conventional use as an optimisation method.
4.2.2 Maximum likelihood estimation

I now outline how to estimate the parameters of the algorithm in Figure 10, using the observed sequence of designs.

For each participant and task (treatment and control condition), I observe data on the order in which designs are submitted \(t\), and the associated creativity \(c\) and cost efficiency \(e\) of each bridge, but do not observe the relative weight on creativity \(\theta\) or the willingness to consider inferior solutions \(d\).

To estimate these unknown parameters, I use the fact that the utility-maximisation algorithm describes a discrete-time inhomogeneous Markov chain with one-step transition probabilities, since next period’s design, conditional on the current design, does not depend on past designs. If relative weight on creativity \(\theta\) or the willingness to consider inferior solutions \(d\) were known, then we could use the utility function \(U = \theta c + (1 - \theta)e\) to partition the data into three groups:

(i) Designs that improve utility: \(U_t \geq U_{t-1}\) (\(n\) observations).

(ii) Designs that do not improve utility \(U_t < U_{t-1}\) but the individual nevertheless continues to work on in the next period (\(m_1\) observations).

(iii) Designs that do not improve utility \(U_t < U_{t-1}\) and that the individual does not continue to work on in the next period (\(m_2\) observations).

The notion of “continues to work on” can be determined via a cutoff rule, where a change in the objective function larger than a specified threshold \((U_{t+1} - U_t > h)\) indicates that the individual chose a substantially different design (both in creativity and cost efficiency) in the next period.

Following the approach of Fleming and Harrington (1977), I can write the maximum likelihood function for the observed data as:

\[
L(U, t, q, d) = Pr(U_t \geq U_{t-1})^n \cdot Pr(U_t < U_{t-1})^{m_1} \prod_{v \in m_1} t_v \frac{\Delta U(v)}{d} \cdot Pr(U_t < U_{t-1})^{m_2} \prod_{s \in m_2} (1 - t_s \frac{\Delta U(s)}{d})
\]

where \(\Delta U(v)\) and \(\Delta U(s)\) denote the difference in the objective function between periods \(t_v\) and \(t_{v-1}\), and \(t_s\) and \(t_{s-1}\) respectively, and the independence (multiplicative terms) comes from the properties of Markov chains (Grimmett and Stirzaker, 2001).

When the relative weight on creativity \(\theta\) or the willingness to consider inferior solutions \(d\) are unknown, they can be jointly estimated via a grid search on \(\theta\). Since creativity \(c\) and cost efficiency \(e\) are known, the relative weight on creativity \(\theta\) determines the utility function \(U\) and thus we can partition the data into the three groups as described by (i)-(iii).

The only remaining unknown is the probability that utility increases, \(Pr(U_t \geq U_{t-1})\). I assume that this probability is constant for a particular individual and task, and denote it as the parameter \(q\).
The log likelihood function can therefore be written as
\[
l(U, t, q, d) = n \cdot \ln(q) + (m_1 + m_2)\ln(1 - q) + \sum_{v \in m_1} \left(- \frac{\Delta U(v)}{d}\right) \ln(t_v) + \sum_{s \in m_2} \ln(1 - t_s - \frac{\Delta U(s)}{d})
\]
where \( q = Pr(U_t \geq U_{t-1}) \).

My grid search method uses the following process:
1. Specify a value for the relative weight on creativity (\( \theta \)).
2. Estimate the parameters (probability that utility increases (\( q \)), and willingness to consider inferior solutions (\( d \)) via maximum likelihood.
3. Evaluate the value of the log likelihood function \( l(U, t, q, d) \) at the optimum.
4. Repeat Steps 1-4 for all values of \( \theta \) in the interval [0, 1], in increments of \( \epsilon \) (I set \( \epsilon = 0.01 \)).

The estimated parameters are those that jointly maximise the log likelihood function.\(^{20}\)

I estimate these parameters at the individual-task level, thus allowing for unrestricted heterogeneity across individuals. Table 1 presents the parameter estimates obtained using this method, where the relative weight on creativity (\( \theta \)) and the probability that the current design increases utility (\( q \)) are bounded by [0, 1], the willingness to consider inferior solutions (\( d \)) is strictly positive (with no upper limit specified), and the threshold used to determine whether an individual continues to work on the design in the next period (\( h \)) is 20. Note that parameter estimates at the individual-task level could only be obtained if the individual submitted three or more designs for that task.

<table>
<thead>
<tr>
<th></th>
<th>Control</th>
<th>Treatment</th>
</tr>
</thead>
<tbody>
<tr>
<td>Relative weight on creativity</td>
<td>0.289</td>
<td>0.327</td>
</tr>
<tr>
<td></td>
<td>(0.346)</td>
<td>(0.363)</td>
</tr>
<tr>
<td>Willingness to consider inferior solutions</td>
<td>257</td>
<td>232</td>
</tr>
<tr>
<td></td>
<td>(380)</td>
<td>(360)</td>
</tr>
<tr>
<td>Probability that current design increases utility</td>
<td>0.597</td>
<td>0.651</td>
</tr>
<tr>
<td></td>
<td>(0.269)</td>
<td>(0.264)</td>
</tr>
<tr>
<td>( N )</td>
<td>273</td>
<td>273</td>
</tr>
</tbody>
</table>

Parameter estimates from maximum likelihood estimation, with standard errors in brackets. The estimation procedure was done separately for each individual and task, under the assumption that each individual’s utility function is a weighted sum of creativity and cost efficiency. Averages for ‘Treatment’ are pooled across all treatment conditions. The relative weight placed on creativity can range from 0 (individual only tries to maximise cost efficiency) to 1 (individual only tries to maximise creativity). The willingness to consider inferior solutions, defined as the continuing to work on a design that gives strictly lower utility than the previous design, must be strictly positive but has no theoretical upper limit.

\(^{20}\)Specifically, the estimated relative weight on creativity (denoted \( \hat{\theta} \)) gives the maximal value of \( l(U, t, q, d) \) across all values of \( \theta \). The estimated willingness to consider inferior solutions and probability that utility increases (denoted \( \hat{d} \) and \( \hat{q} \) respectively) are the values that jointly maximise \( l(U, t, q, d) \) with \( \hat{\theta} \).
5 Results

5.1 Descriptive statistics

Table 1 summarises the key features of my sample (a full summary table is in Table 6, Appendix C). 37% of participants were female, 11% were left handed, and the average age was 26-30 years old. On average, participants had previously tried 5 of the 9 types of similar activities surveyed, and had studied a relevant subject in university or high school (such as engineering, physics, or architecture) for one to two years. The average participant submitted 5 designs per task, and half of these designs were structurally stable. Pooled across all individuals and tasks, the average edit distance of the most creative bridge was 19 (subject to satisfying the budget and structural stability criteria), and the cheapest stable bridge cost $347,000 on average (see Figure 22 in Appendix C for the joint and marginal distributions of creativity and cost efficiency).

Table 2. Summary statistics: Demographics and characteristics of output

<table>
<thead>
<tr>
<th>Variable</th>
<th>Mean</th>
<th>SD</th>
</tr>
</thead>
<tbody>
<tr>
<td>Female</td>
<td>0.37</td>
<td>0.48</td>
</tr>
<tr>
<td>Age (5-year bands)</td>
<td>2.13</td>
<td>1.51</td>
</tr>
<tr>
<td>Income (personal annual before-tax, in £000s)</td>
<td>18.5</td>
<td>18.5</td>
</tr>
<tr>
<td>Left handed</td>
<td>0.11</td>
<td>0.32</td>
</tr>
<tr>
<td>Number of similar activities done before (sum)</td>
<td>5.60</td>
<td>2.04</td>
</tr>
<tr>
<td>Number of relevant subjects studied (weighted sum)</td>
<td>2.52</td>
<td>1.71</td>
</tr>
<tr>
<td>Number of designs (per task)</td>
<td>5.40</td>
<td>7.08</td>
</tr>
<tr>
<td>Stable designs (proportion, per task)</td>
<td>0.51</td>
<td>0.42</td>
</tr>
<tr>
<td>Creativity (max. edit distance, per individual and task)</td>
<td>19.4</td>
<td>12.5</td>
</tr>
<tr>
<td>Cost (cheapest stable bridge ($000s), per individual and task)</td>
<td>347.2</td>
<td>40.2</td>
</tr>
</tbody>
</table>

Summary statistics (mean and standard deviation (SD)) for experiment participants (N=550). Gender is an indicator variable that equals 1 if female. Age is measured in 5-year bands, where 1 equals 25 and under, 2 equals 26-30, and so on until 8 (56 or older). Income refers to personal annual before-tax income, excluding any benefits received from the government or receipts from other sources. Left-handed is an indicator that equals 1 if the participant normally uses their left hand for writing, where ambidextrous individuals (0.65% of participants) are counted as left handed. Number of relevant subjects is a weighted sum, where subjects that were studied for more than one secondary school year/one university term were given double the weight compared to subjects that were studied for one secondary school year/one university term or less. Relevant subject categories were physical sciences, engineering/computer science, and design. Structurally stable designs are bridges that can bear the weight of a standard-size truck. Creativity was measured using graph edit distance, defined as the minimal number of operations required to transform one bridge design into a pre-specified reference bridge design. Cost refers to the cost of materials used to produce the bridge.
5.2 Mean treatment effects

First, I present results of the mean effect of incentives on the creativity of bridge designs, focusing on the most creative (highest edit distance) design submitted by each participant in the control and treatment condition. Only designs that satisfied the budget constraint and were structurally stable were considered.

Panel (a) of Figure 11 shows the within-person difference in creativity (treatment minus control) for each of the six treatment conditions, where positive numbers indicate that an individual produced more-creative output in the treatment condition compared to the control condition. The first two bars in panel (a) show that creativity increased by 3-4 units when directly incentivised, and that both of these effects are significant at the 10% level ($p = 0.057$ for creativity (high) and $p = 0.076$ for creativity (low)). The middle two bars show that participant’s most creative design did not significantly differ in creativity when they were incentivised for cost efficiency, an unsurprising result. The rightmost two bars show that giving participants extra time but no performance-related incentives results in similar effect sizes compared to creativity-specific incentives, though there was a wider variation in responses, so the mean effects are not statistically significant ($p = 0.416$ for extra time (high) and $p = 0.143$ for extra time (low)). Figure 23 in Appendix C shows the distribution of mean differences for each treatment.

Due to aspects of the experimental design (randomisation into treatments, and random order of treatment and control), it is unlikely that these effects are due to selection into particular incentive schemes, or due to learning. Table 7 in Appendix C presents evidence of demographic balance across groups: ANOVA tests for the differences in means across groups fail to reject the null hypothesis at any conventional significance level.\footnote{One exception is task enjoyment, but this result is entirely driven by a higher mean in one treatment (efficiency (high)); removing this treatment from the ANOVA test results in an F-statistic of 1.76 ($p = 0.135$).}

One potential explanation for the significant direct effect on creativity is that participants compromised on cost efficiency in order to produce a more creative bridge. To explore this issue, I compare the associated cost of each participant’s most creative bridge in the treatment and control condition. I define cost efficiency as the amount of budget remaining, so higher numbers indicate cheaper bridges, and a positive difference in cost efficiency indicates that the participant’s most creative bridge in the treatment condition was cheaper than their most creative bridge in the control condition.

Panel (b) shows there is no evidence of compromising on cost efficiency: on average, direct incentives on creativity had an insignificant impact on cost efficiency (both statistically and in terms of effect size).\footnote{Since bridges cost $340,000 on average, a $10,000 decrease in cost efficiency corresponds to a 3% increase in cost, which is a relatively small magnitude.} In other words, a participant’s most creative bridge in the treatment...
condition was not substantially more expensive than their most creative bridge in the control condition.

**Figure 11.** Direct and indirect effects of pay-for-performance incentives: Creativity

![Bar Chart](chart.png)

This figure shows that pay-for-performance incentives on creativity have positive direct effects and insignificant indirect effects. In both panels, the bars show the average within-person difference in the variable of interest (treatment minus control): creativity of the most creative design (panel (a)), and the associated cost efficiency of that design (panel (b)). Creativity is measured using edit distance, defined as the minimal number of operations required to transform one bridge design into the pre-specified reference bridge design. Cost efficiency ($000s) is measured as the amount of budget remaining ($400000 minus the cost of production). In both panels, the mean difference (treatment minus control) is reported above each bar, with standard errors in brackets. Under the null hypothesis that treatments have no effect on the mean outcome of interest, differences that are statistically significant at the 10% level ($p < 0.1$) are shaded in grey.

Next, I present results of the mean effect of incentives on the cost efficiency of bridge designs, focusing on the cheapest (“most cost-effective”) design submitted by each participant in the control and treatment condition. Only designs that satisfied the budget constraint and were structurally stable were considered.

Panel (a) of Figure 12 shows the within-person difference in cost efficiency (treatment minus control) for each of the six treatment conditions, where positive numbers indicate that an individual produced cheaper output in the treatment condition compared to the control condition. The middle two bars in panel (a) show that on average, incentives on cost efficiency have positive but insignificant direct effects: an individual’s design was on average $240 cheaper for efficiency (high) ($p = 0.340$), and $6190 cheaper for efficiency (low) ($p = 0.483$). The rightmost two bars show that giving participants extra time but no performance-related incentives results in similar effects (both in size and significance) compared to cost efficiency incentives. Figure 24 in Appendix C shows the distribution of mean differences for each treatment.

The middle two bars in panel (b) shows there is no evidence of compromising on cost efficiency: on average, direct incentives on cost had an insignificant impact on creativity (both statistically and in terms of effect size).\(^{23}\) In other words, a participant’s cheapest bridge

\(^{23}\)Since bridges cost $340,000 on average, a $10,000 decrease in cost efficiency corresponds to a 3% increase in cost, which is a relatively small magnitude.
in the treatment condition was not substantially more or less creative than their cheapest bridge in the control condition.

**Figure 12.** Direct and indirect effects of pay-for-performance incentives: Cost efficiency

![Bar charts showing direct and indirect effects of pay-for-performance incentives on cost efficiency and creativity.](image)

This figure shows that pay-for-performance incentives on cost efficiency have insignificant direct and indirect effects (both statistically and in terms of effect size). In both panels, the bars show the average within-person difference in the variable of interest (treatment minus control): cost efficiency of the cheapest design (panel (a)), and the associated creativity of that design (panel (b)). Creativity is measured using edit distance, defined as the minimal number of operations required to transform one bridge design into the pre-specified reference bridge design. Cost efficiency ($000s) is measured as the amount of budget remaining ($400000 minus the cost of production). In both panels, the mean difference (treatment minus control) is reported above each bar, with standard errors in brackets. Under the null hypothesis that treatments have no effect on the mean outcome of interest, differences that are statistically significant at the 10% level ($p < 0.1$) are shaded in grey.

### 5.3 Possible explanations

#### #1. Incentives increase effort

I consider whether the significant increase in creativity for the creativity-specific treatments is due to greater effort, either in terms of submitting a greater number of total designs or submitting a greater number of usable (structurally stable) designs. For participants who had goal-specific incentives, only stable bridges could be considered for the bonus, so the number of stable designs submitted is a proxy for the quality of submissions (independent of the originality of submissions).

Figure 13 shows the within-person differences for both of these measures. On average, participants with creativity-specific incentives did not submit a greater number of total designs, or a greater number of stable designs. The average number of designs increases for efficiency (high), and in three of the four non-creativity-focused treatments, the number of stable designs increased, though the effect size is small (0.3-0.8 designs on average).
**Figure 13.** Mean within-person differences in effort (total number of designs submitted, and number of stable designs submitted)

This figure shows how two measures of effort (quantity and quality) differ in the treatment and control condition. In both panels, the bars show the average within-person difference in the variable of interest (treatment minus control): the total number of designs submitted per task by each participant (panel (a)), and the number of structurally stable designs submitted per task by each participant (panel (b)). For participants who had goal-specific incentives, only stable bridges could be considered for the bonus, so the number of stable designs submitted is a proxy for the quality of submissions. In both panels, the mean difference (treatment minus control) is reported above each bar, with standard errors in brackets. Under the null hypothesis that treatments have no effect on the mean outcome of interest, differences that are statistically significant at the 10% level ($p < 0.1$) are shaded in grey.

**2. Incentives affect the degree of exploration**

Another potential explanation is that participants facing creativity-specific incentives searched more broadly, trying a wider variety of designs in an attempt to find better designs. To measure this concept, I look at variations in one important and salient aspect of the design: creativity, and calculate the standard deviation in the edit distance of designs submitted by each participant. If participants explore by testing designs that have a larger variation in creativity, we would expect the standard deviation to increase. Figure 14 shows the mean within-person difference in the standard deviation of edit distance, where positive numbers indicate a larger standard deviation in the treatment relative to the control condition. On average, the designs of participants facing the creativity (high) incentives had a larger variation in creativity (measured by edit distance) in the treatment condition, though the difference is not statistically significant ($p = 0.22$). Conversely, the designs of participants incentivised for cost efficiency had smaller variations in creativity (measured by edit distance) in the treatment condition ($p = 0.12$ for efficiency (high) and $p = 0.07$ for efficiency (low)).
**Figure 14.** Mean within-person difference: Standard deviation of edit distance in designs submitted (treatment minus control)

This figure shows that participants incentivised for cost efficiency explored a narrower variety of designs in the treatment condition. The bars show the average within-person difference in the standard deviation of edit distance in designs submitted (treatment minus control), with standard errors in brackets. Creativity is measured using edit distance, defined as the minimal number of operations required to transform one bridge design into the pre-specified reference bridge design. Under the null hypothesis that treatments have no effect on the mean outcome of interest, differences that are statistically significant at the 10% level ($p < 0.1$) are shaded in grey.

#3 Heterogeneous responses to treatment, and selection effects

Next, I explore two possible explanations: firstly, it is possible that the null mean effect masks heterogeneity in response to treatment. Previous studies have shown that demographic variables and personality characteristics can be significantly correlated with performance in creative tasks (Erat and Gneezy 2016, Charness and Grieco 2019). It is therefore possible that the mean within-person effects observed in Figures 11 and 12 are driven by participants who satisfy certain characteristics. Secondly, since bridges had to satisfy the budget constraint and the stability requirement to be considered “useful”, it is possible for participants to produce zero useful output in a particular task (all bridges fail the load test, cost more than the given budget, or both). If there are characteristics that significantly affect the probability of producing useful output and performance in the task, for example, skill in bridge design, then the mean within-person effects could be upward-biased.

To examine these issues, I use a Heckman regression to jointly estimate the effect of outcomes (creativity of most creative bridge, cost efficiency of cheapest bridge) and the probability of producing useful output in both tasks (observing complete data for an individual). Regressions are pooled across all treatments and include demographic variables such as age and gender, exposure to similar tasks or relevant knowledge, treatment-specific indicator variables (with extra-time-low as the base category), measures of personality traits (scores on the Big-5 personality test, grit score, willingness to take risks), and changes in characteristics of output produced in the treatment condition (change in the number of designs submitted, change in the percentage of designs that are structurally stable, and change in standard deviation of creativity). A full summary of these variables is in Table 6 (Appendix C).
Panel (a) of Figure 15 show estimates from the outcome equation, where the within-person change in creativity (treatment minus control) is the dependent variable. Estimates are presented as the effect of a 1-standard-deviation increase in the independent variable, with 95% confidence intervals. Across all treatments, larger positive treatment effects are associated with participants who tried a wider variety of designs compared to the control condition, and were more intrinsically motivated. The significant negative coefficient on initial creativity (most creative design in the control condition) indicates that the mean within-person effects observed for creativity could be driven by participants in the lower end of the distribution; in other words, participants who had a lower baseline performance had a larger increase in creativity when treated.

Panel (b) of Figure 15 shows estimates from the outcome equation, where the within-person change in cost efficiency (treatment minus control) is the dependent variable. Again, participants who had a lower baseline performance (produced more expensive designs) had a larger increase in cost efficiency in the treatment condition. This result may be due to the stability criterion, which makes it more difficult to improve cost efficiency when a bridge is already cheap.

Figure 16 shows coefficients from the selection equation, where the dependent variable takes the value 1 if individual-level data for both treatment and control conditions are nonmissing. I am more likely to observe complete data for individuals who are more intrinsically motivated, and whose reported problem-solving approach is more experimental (trial and error) rather than systematic. I am less likely to observe complete data for females, suggesting gender differences in overall performance on the task.

One point to note is that the creativity goal was less transparent than the cost efficiency goal, because participants could observe the cost in real time with complete certainty, but could not receive any feedback during the experiment about how creative their designs were. Participants may have different interpretations of the creativity goal, which in turn may affect performance (as evaluated by my measure). To investigate whether performance is correlated with participants’ notion of visual dissimilarity, I include a dummy variable that equals 1 if the participant correctly ranked the 4 bridges given in the exit survey. At the 5% level, I fail to reject the null hypothesis that performance (both producing useful output and outcomes conditional on producing useful output) is significantly correlated with participants’ notion of visual dissimilarity.
Panels (a) and (b) show the change in the dependent variable associated with a one-standard-deviation increase in the independent variable. All point estimates are shown with 95% confidence intervals, based on robust standard errors. All regressions are pooled across treatments and include treatment-specific indicator variables (with extra-time-low as the base category), as well as a constant term.
#4 Incentives affect the creative process

Previous literature has found that financial incentives can affect the size of the solution set that an individual considers, and in some settings discourage exploration of new approaches (Ederer and Manso 2013, Englemaier et al 2018). To further explore how goal-specific incentives affect creative process in my setting, I look at how the maximum-likelihood parameters (relative weight on creativity and willingness to consider inferior solutions) change in response to treatment.

Figure 17 shows the mean within-person differences in these parameters (treatment minus control). With the exception of extra time (high), incentives did not significantly affect the relative weight placed on the creativity goal (vs. the cost efficiency goal) on average. However, the average willingness to consider inferior solutions significantly increased under creativity (high) \((p = 0.02)\) and extra time (high) \((p = 0.05)\), and decreased for efficiency (high) \((p = 0.003)\). To provide an intuitive interpretation of the effect size, an increase of 160 in the willingness to consider inferior solutions, evaluated at the sample mean, roughly corresponds to a 10 percentage-point increase in the probability of continuing to work on an inferior design, and a decrease of 209 roughly corresponds to a 58 percentage-point decrease in the probability of continuing to work on an inferior design.\(^{24}\)

\(^{24}\)The probability of continuing to work on an inferior design is defined as \( t^{-\frac{\triangle Utility}{Willingness}} \), where \( t \) is the time period. For the values \( t = 5, \triangle Utility = 50 \), and a willingness of 250, an increase of 160 changes the
Figure 17. Mean within-person difference in maximum likelihood parameters

The bars in both panels show mean within-person differences (treatment minus control) in the variable of interest. Panel (a) shows the change in the relative weight an individual places on creativity compared to cost efficiency in their objective function, which ranges from 0 (solely focusing on maximising cost efficiency) to 1 (solely focusing on maximising creativity). Panel (b) shows the change in the willingness to consider inferior solutions, which is proportional to the probability of continuing to work on a design that is less optimal than the previous design. Standard errors are reported in brackets. Under the null hypothesis that treatments have no effect on the mean outcome of interest, differences that are statistically significant at the 10% level ($p < 0.1$) are shaded in grey.

These mean effects mask substantial variation. Figure 25 (Appendix C) shows that the distributions of both variables have large ranges and outliers (values that are more than 1.5 standard deviations away from the first and third quartiles). Figure 26 (Appendix C) explores the correlations between the willingness to consider inferior solutions and individual characteristics. Across all treatments, left-handed individuals had a larger increase in the willingness to consider inferior solutions, and individuals who had a high baseline willingness had a smaller increase, ceteris paribus.

To examine the relationship between relative weight on creativity and the willingness to consider inferior solutions affect performance in the creativity and cost efficiency goals, I re-run the regressions in Figure 15, with these variables as additional regressors. To test for the possibility of non-linear effects, I include squared terms of both variables. Table 3 shows the estimated coefficients for these variables. Across all treatments, the willingness to consider inferior solutions (both the linear and squared terms) is significantly correlated with performance in the creativity goal at the 10% level, but not with performance in the cost efficiency goal. In both regressions there is a negative quadratic relationship between this variable and performance, suggesting that intermediate levels are optimal. The relative weight on creativity is not significantly correlated with performance in the creativity goal, but the squared term is significantly correlated with performance in the cost efficiency goal ($p = 0.049$).

---

| Probability of continuing to work on an inferior design | From 0.72 to 0.82 (a 10 percentage-point increase), and a decrease of 209 changes the probability from 0.72 to 0.14 (a 58 percentage-point decrease). | 32 |
Table 3. Correlation between features of the creative process and outcomes (creativity of the most creative bridge, and cost efficiency of the cheapest bridge)

<table>
<thead>
<tr>
<th></th>
<th>Creativity</th>
<th>Cost efficiency ($000s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Relative weight on creativity</td>
<td>-5.55</td>
<td>25.58</td>
</tr>
<tr>
<td></td>
<td>(5.43)</td>
<td>(21.867)</td>
</tr>
<tr>
<td>Relative weight on creativity²</td>
<td>7.24</td>
<td>-46.40</td>
</tr>
<tr>
<td></td>
<td>(5.90)</td>
<td>(23.59)</td>
</tr>
<tr>
<td>Willingness to consider inferior solutions (00s)</td>
<td>1.72</td>
<td>5.20</td>
</tr>
<tr>
<td></td>
<td>(0.98)</td>
<td>(6.35)</td>
</tr>
<tr>
<td>Willingness to consider inferior solutions (00s)²</td>
<td>-0.0016</td>
<td>-0.0056</td>
</tr>
<tr>
<td></td>
<td>(0.00094)</td>
<td>(0.0060)</td>
</tr>
</tbody>
</table>

This table shows estimated coefficients on the maximum likelihood parameters, from Heckman regressions that are pooled across all treatments, and include demographic variables such as age and gender, exposure to similar tasks or relevant knowledge, treatment-specific indicator variables (with extra-time-low as the base category), measures of personality traits (scores on the Big-5 personality test, grit score, willingness to take risks), and changes in characteristics of output produced in the treatment condition (change in the number of designs submitted, change in the percentage of designs that are structurally stable, and change in standard deviation of creativity). Creativity is measured using edit distance, defined as the minimal number of operations required to transform one bridge design into the pre-specified reference bridge design. Cost efficiency ($000s) is measured as the amount of budget remaining ($400000 minus the cost of production). Column titles indicate the dependent variable. Robust standard errors, clustered at the individual level, are reported in brackets.

6 Discussion

This paper has shown that creativity-specific incentives have sizeable positive direct effects. One explanation for this effect is that these incentives change the way that individuals develop and refine output, which in turn leads to better outcomes. I find evidence that individuals facing high-powered incentives on creativity increase their level of exploration, both in terms of the variety of designs attempted and the willingness to consider inferior solutions, and that the level of exploration is positively and significantly correlated with creative task performance.

One surprising result is the absence of negative spillover effects in either direction: under incentives specifically on creativity, participants’ most-creative designs were not significantly more expensive, and under incentives specifically on cost efficiency, participants’ cheapest designs were not significantly less creative. The results in Table 3 suggest one possible explanation: the creativity goal is more sensitive to the level of exploration than the cost efficiency goal. In other words, relative to the range of values that creativity and cost can take, the “returns to exploration” for cost efficiency are much smaller than those for creativity.

Taken at the mean of the respective dependent variables, a 100-unit increase in the willingness to experiment (from 200 to 300) is associated with a 5% increase in creativity (1-unit increase in edit distance, relative to a mean of 20) that is significant at the 10% level ($p = 0.084$), but only a 0.007% increase in cost efficiency ($2370 relative to a mean of $350,000). Therefore,
if individuals increase their degree of exploration in response to the pay-for-performance incentives, it is possible to have a large response in creativity without an accompanying decrease in cost efficiency.

Another point for discussion is the comparable mean effect of the flat-payment extra-time treatments. A longstanding debate in the literature is whether or not incentives are necessary to stimulate creativity (Charness and Grieco, 2019). From a firm’s perspective, if we can achieve the same effects as financial incentives simply by giving individuals additional time to develop their ideas, then the latter might be a cheaper and more cost effective way to produce creative output.

I compare the cost (to the researcher) of the six treatments in my experiment, using the control condition as a baseline. Panel (a) of Figure 18 shows the average cost of each treatment, measured in pounds (£) per person. According to this measure, the extra-time treatments are the most expensive because they involve a flat payment regardless of performance, whereas goal-specific incentives involve a relatively low base payment (£0.50 for “high” treatments and £1.50 for “low” treatments). Creativity (high) had the lowest average cost (£0.99).

To measure cost effectiveness, I calculate the units of creativity per dollar (or pound) spent, by dividing each individual’s maximum creativity score by their payment. For example, if an individual’s most creative bridge had an edit distance of 20 and they were paid £5.00 for that task, that would correspond to 4 units of creativity per pound spent. Individuals who failed to produce useful output in a particular task (submitted bridges that were too expensive, not stable, or both) would have a value of 0.

Panel (b) of Figure 18 shows the average units of creativity per pound spent, with 95% confidence intervals. By this measure, creativity (high) is the most cost-effective treatment: the average units of creativity per £ is significantly higher than the baseline ($p < 0.001$) and higher than extra time (low) at the 10% level ($p = 0.064$).

Figure 18. Comparison of cost (a) and cost effectiveness (b)
My results highlight some promising avenues for future study. Both the reduced-form and structural estimation results suggest that in this two-goal setting, intermediate levels of exploration are optimal: exploring too much could result in failure to meet one or more criteria, but exploring too little could result in failure to improve on the outcomes of interest. However, the most effective way for firms to encourage this optimal level of exploration is still an open question. My study provides some evidence that direct incentives on creativity can significantly increase individuals’ willingness to consider inferior solutions, but it is possible that some other form of treatment (besides those considered in this paper) may induce individuals to explore more, and also be more cost effective than pay-for-performance incentives. In single-goal settings, non-financial incentives such as peer ranking were found to facilitate creativity in “closed” tasks similar to the one I use in this experiment (Charness and Grieco 2019). Whether these results extend to multi-goal settings is a topic for future research.

7 Conclusion

Creativity plays an important role in modern economies and is highly valued by firms. In order to foster creativity, firms may choose to use financial incentives such as pay-for-performance or tournaments. While there is some experimental and field evidence that these types of incentives can increase creativity if that is the worker’s only objective, it is unclear whether these findings extend to settings that characterise many business environments, specifically where the production of creative output involves an explicit financial cost.

Using a lab experiment and a novel task that can vary in two dimensions (creativity and cost efficiency), I investigate the direct and indirect effects of goal-specific pay-for-performance incentives on the production of creative output, focusing both on characteristics of the output produced and the process of developing it. Besides providing an experimental test of how pay-for-performance incentives affect creative outcomes in two-goal settings, including the potential for spillover effects, I also make two methodological contributions to the literature. First, I employ the graph edit distance metric to measure creativity cleanly and precisely in the context of my experiment. Second, I develop a utility-maximisation framework that describes how individuals search the two-dimensional space for optimal solutions. This framework enables me to estimate the willingness to experiment and the relative weight on each goal from the observed sequence of outputs at the individual level, allowing for unrestricted heterogeneity across individuals.

I find that pay-for-performance incentives on creativity have sizeable positive direct effects without significant indirect effects on cost efficiency. The degree of exploration, measured in terms of the variety of designs produced and the willingness to consider inferior solutions, is crucial for explaining both the average positive direct effect and null indirect effect. Under high-powered incentives on creativity, individuals increase their level of exploration, which in
turn improves their creative task performance. The absence of significant spillover effects is partly due to the larger relative effect size of exploratory behaviour on creativity compared to cost efficiency. Both the reduced-form and structural estimation results suggest that in two-goal settings, intermediate levels of exploration are optimal.

Overall, my results provide further evidence that the creative process is important for understanding the relationship between incentives and outcomes, and that exploration (trying a variety of solutions and being willing to consider inferior solutions) plays a crucial role in creative task performance. Direct financial incentives on creativity may still be effective in settings where the production of creative output involves an explicit financial cost, provided that performance in each goal is positively associated with the degree of exploration.
Appendix A: Experiment instructions and questionnaire

Control condition

Task 1

Open the file ‘arch_bridge.bdc’.

Your goal is to make a bridge that a truck can successfully drive over, and also:
- Contains an arch (an inverted ‘u’ shape),
- Costs $400,000 or less,
- Looks as different from the ‘standard’ arch bridge design as possible.

(Note: ‘Different’ is taken in the literal sense, so bridges that are asymmetrical or irregularly shaped in any way are also considered different.)

You have 20 minutes to work on this task. Then stop, save all your bridges, and upload them to the survey as a .zip file (see next page for instructions on how to do this.

You can make as many bridges as you want in the time allocated (and not just one bridge).

You will be paid £2.50 regardless of your performance on this task.

Examples of designs and their similarity to the standard design are shown below.
Creativity (high)

Task 2

Open the file ‘truss_bridge.bdc’.

Your goal is to make a bridge that a truck can successfully drive over, and also:
- Does not contain an arch (inverted ‘u’ shape),
- Costs $400,000 or less,
- Looks as different from the ‘standard’ non-arch bridge design as possible.

(Note: ‘Different’ is taken in the literal sense, so bridges that are asymmetrical or irregularly shaped in any way are also considered different.)

You have 20 minutes to work on this task. Then stop, save all your bridges, and upload them to the survey as a .zip file (see next page for instructions on how to do this.)

You can make as many bridges as you want in the time allocated (and not just one bridge).

You will be paid £0.50 regardless of your performance on this task.

In addition, you will be paid a bonus ranging from £0.01 to £4, according to how different your bridge is from the standard bridge, in terms of design.

(Note: the bonus has been calibrated according to the parameters of the bridge template so that the average bonus will be around £2).

If you build more than one successful bridge, the bonus will be determined according to your most different bridge.

Examples of designs and their similarity to the standard design are shown below.

Example #1: Quite similar  
Bonus: £0.11

Example #2: Less similar  
Bonus: £0.85
Creativity (low)

Task 2

Open the file ‘truss_bridge.bdc’.

Your goal is to make a bridge that a truck can successfully drive over, and also:
- Does not contain an arch (inverted ‘u’ shape),
- Costs $400,000 or less,
- Looks as different from the ‘standard’ non-arch bridge design as possible.

(Note: ‘Different’ is taken in the literal sense, so bridges that are asymmetrical or irregularly shaped in any way are also considered different.)

You have 20 minutes to work on this task. Then stop, save all your bridges, and upload them to the survey as a .zip file (see next page for instructions on how to do this.)

You can make as many bridges as you want in the time allocated (and not just one bridge).

You will be paid £0.50 regardless of your performance on this task.

In addition, you will be paid a bonus ranging from £1 to £3, according to how different your bridge is from the standard bridge, in terms of design.

(Note: the bonus has been calibrated according to the parameters of the bridge template so that the average bonus will be around £2).

If you build more than one successful bridge, the bonus will be determined according to your most different bridge.

Examples of designs and their similarity to the standard design are shown below.

Example #1: Quite similar
Bonus: £1.05

Example #2: Less similar
Bonus: £1.43
Cost efficiency (high)

Task 2

Open the file ‘truss_bridge.bdc’.

Your goal is to make a bridge that a truck can successfully drive over, and also:
- Does not contain an arch (inverted ‘u’ shape),
- Costs $400,000 or less,
- Looks as different from the ‘standard’ non-arch bridge design as possible.

(Note: ‘Different’ is taken in the literal sense, so bridges that are asymmetrical or irregularly shaped in any way are also considered different.)

You have 20 minutes to work on this task. Then stop, save all your bridges, and upload them to the survey as a .zip file (see next page for instructions on how to do this.)

You can make as many bridges as you want in the time allocated (and not just one bridge).

You will be paid £0.50 regardless of your performance on this task.

In addition, you will be paid a bonus ranging from £0.01 to £4, according to how cheap your bridge is (proportional to $400,000 minus the cost of your bridge).

(Note: the bonus has been calibrated according to the parameters of the bridge template so that the average bonus will be around £2).

If you build more than one successful bridge, the bonus will be determined according to your cheapest bridge.

Examples of designs and their similarity to the standard design are shown below.

Example #1: Quite similar
Cost: $316,470
Bonus: £1.67

Example #2: Less similar
Cost: $323,683
Bonus: £1.53
Task 2

Open the file ‘truss_bridge.bdc’.

Your goal is to make a bridge that a truck can successfully drive over, and also:
- Does not contain an arch (inverted ‘u’ shape),
- Costs $400,000 or less,
- Looks as different from the ‘standard’ non-arch bridge design as possible.

(Note: ‘Different’ is taken in the literal sense, so bridges that are asymmetrical or irregularly shaped in any way are also considered different.)

You have 20 minutes to work on this task. Then stop, save all your bridges, and upload them to the survey as a .zip file (see next page for instructions on how to do this.)

You can make as many bridges as you want in the time allocated (and not just one bridge).

You will be paid £0.50 regardless of your performance on this task.

In addition, you will be paid a bonus ranging from £1 to £3, according to how cheap your bridge is (proportional to $400,000 minus the cost of your bridge).

(Note: the bonus has been calibrated according to the parameters of the bridge template so that the average bonus will be around £2).

If you build more than one successful bridge, the bonus will be determined according to your cheapest bridge.

Examples of designs and their similarity to the standard design are shown below.

Example #1: Quite similar
Cost: $316,470
Bonus: £1.84

Example #2: Less similar
Cost: $323,683
Bonus: £1.76
Extra time (high)

Task 2

Open the file ‘truss_bridge.bdc’.

Your goal is to make a bridge that a truck can successfully drive over, and also:
- Does not contain an arch (inverted ‘u’ shape),
- Costs $400,000 or less,
- Looks as different from the ‘standard’ non-arch bridge design as possible.

(Note: ‘Different’ is taken in the literal sense, so bridges that are asymmetrical or irregularly shaped in any way are also considered different.)

‘Standard’ non-arch design
Cost = $292,877

You have 40 minutes to work on this task. Then stop, save all your bridges, and upload them to the survey as a .zip file (see next page for instructions on how to do this.)

You can make as many bridges as you want in the time allocated (and not just one bridge).

You will be paid the same hourly rate as before (£2.50 per 20 minutes = £5 total) regardless of your performance on this task.

Examples of designs and their similarity to the standard design are shown below.
Extra time (low)

Task 2

Open the file ‘truss_bridge.bdc’.

Your goal is to make a bridge that a truck can successfully drive over, and also:
• Does not contain an arch (inverted ‘u’ shape),
• Costs $400,000 or less,
• Looks as different from the ‘standard’ non-arch bridge design as possible.

(Note: ‘Different’ is taken in the literal sense, so bridges that are asymmetrical or irregularly shaped in any way are also considered different.)

You have 40 minutes to work on this task. Then stop, save all your bridges, and upload them to the survey as a .zip file (see next page for instructions on how to do this.)

You can make as many bridges as you want in the time allocated (and not just one bridge).

You will be paid the same amount as before (£2.50) regardless of your performance on this task.

Examples of designs and their similarity to the standard design are shown below.

Example #1: Quite similar
Example #2: Less similar
Questionnaire

The following questions have no right or wrong answers; please answer them as honestly as you can. All of the questions are optional and you will be paid regardless of your answers.

1. Below is a number of statements, each of which starts with "I see myself as someone who...". For each statement, indicate how much you agree with this (1 = Strongly disagree, 7 = Strongly agree).

I see myself as someone who ...
- Worries a lot
- Gets nervous easily
- Remains calm in tense situations
- Is talkative
- Is outgoing, sociable
- Is reserved
- Is original, comes up with new ideas
- Values artistic, aesthetic experiences
- Has an active imagination
- Is sometimes rude to others
- Has a forgiving nature
- Is considerate and kind to almost everyone
- Does a thorough job
- Tends to be lazy
- Does things efficiently

2. For each of the following statements, indicate how well it describes you (1 = Not like me at all, 5 = Very much like me):

- New ideas and projects sometimes distract me from previous ones.
- Setbacks don’t discourage me.
- I have been obsessed with a certain idea or project for a short time but later lost interest.
- I am a hard worker.
- I often set a goal but later choose to pursue a different one.
- I have difficulty maintaining my focus on projects that take more than a few months to complete.
- I finish whatever I begin.
- I am diligent.
3. For each of the following pairs of statements, select the number that best describes where you are on the scale, where 1 means complete agreement with statement A, and 7 means complete agreement with statement B.

A.1 Planning is essential for me to be creative. I often have detailed sketches for what I am going to do before I do anything.
B.1 Planning is not important for me to be creative. I rarely have detailed sketches for what I am going to do before I do anything.

A.2 I view working creatively as the systematic execution of a plan; I work easily and swiftly.
B.2 I view working creatively as mainly trial and error; I make choices, change them, and react to my changes.

A.3 I have a discontinuous creative career. Once I master one idea or topic, I move on to the next.
B.3 I am a perfectionist who is constantly searching. I am frustrated by my inability to achieve my goals.

A.4 I am finished working creatively when I complete my preconceived plan.
B.4 I am finished working only after inspecting and judging my work.

A.5 When working creatively, I precisely state my goals before beginning, either as an image or an exact procedure.
B.5 When I am working creatively, my goals are imprecise. Having imprecise goals leads me to use a tentative procedure.

A.6 I work creatively to produce something that achieves a purpose.
B.6 I work creatively to search for and discover the meaning of my work.

A.7 My innovation appears suddenly. My new ideas are very different from my old ideas.
B.7 My innovation appears through pursuing one image at a time. My new ideas tend to be different versions of the same thing.

4. How do you see yourself: Are you generally a person who is fully prepared to take risks or do you try to avoid taking risks? Please tick a box on the scale, where the value 0 means: “not at all willing to take risks” and the value 10 means: “very willing to take risks”.

5. How much did you enjoy the creative task (building a bridge)? Please tick a box on the scale, where the value 1 means: ‘did not enjoy the task at all’ and the value 5 means: ‘completely enjoyed the task’.

6. Overall, how well do you think did you on the creative task, relative to other people who did the same task? (1 = Much worse, 5 = Much better)

7. Rank the following bridges according to how different you think they are from the standard bridge (click and drag to put the most different at the top, least different at the bottom). A picture of the standard bridge is at the bottom of the page for your reference.
8. List as many alternative uses for a paperclip as you can think of in three minutes (the page will automatically move to the next question once the time is up).

Demographic questions
Please provide the following information about yourself.

8. What gender best describes you? (Please tick one box.)
   □ Male □ Female

9. What is your age (in years)? (Please tick one box.)
   □ 25 or less □ 26-30 □ 31-35 □ 36-40 □ 41-45 □ 46-50 □ 51-55 □ 56 or more

10. Which hand do you normally use for writing? (Please tick one box.)
    □ Left □ Right □ Both (ambidextrous)

11. Counting all your sources of income, what is your annual before-tax income, rounded to the nearest £5,000 (excluding any benefits received from the government or receipts from other sources)?

11. Did you study the following types of subjects in the last four years of school and/or during university/college (if applicable)? (0 = Never, 1 = Yes, for one school year/one university term or less, 2 = Yes, for more than one school year/one university term)
    - Humanities (e.g. history, geography)
    - Social sciences (e.g. economics, psychology)
    - Biological sciences (e.g. biology, zoology)
    - Physical sciences (e.g. physics, chemistry)
    - Engineering or computer science
    - Design (e.g. art, architecture, fashion)
- Performing arts (e.g. music, drama, film)
12. Before today, have you ever... (0 = Never, 1 = Yes, but only for school/university course work, 2 = Once or twice (excluding school/university course work), 3 = 3-5 times (excluding school/university course work), 4 = More than 5 times (excluding school/university course work))
- Created an original artwork (e.g. painting, sculpture)
- Written an original poem, novel, short story, or graphic novel/webtoon
- Made a craft item (e.g. leather craft, sewing, woodwork)
- Composed an original song
- Directed or acted in a play, film/video, musical, or dance performance
- Wrote an original computer program (e.g. for an app/website, to analyse data)
- Completed a jigsaw puzzle (300 or more pieces) or played games that involved problem-solving (e.g. escape rooms, crosswords)
- Played simulation games that involve design e.g. Sims
- Played a bridge building simulation

**Appendix B: Methodology for calculating edit distance**

I use adjacency matrices to calculate the exact graph edit distance between two bridges. Figure 19 illustrates the methodology. Each bridge can be written in matrix form, where columns represent nodes and entry \(i,j\) takes the value 1 if node \(i\) is connected to node \(j\).

The total edit distance is the sum of the number of nodes added, the number of beams added, and the number of beams moved. These operations can be counted by comparing the adjacency matrices of both bridges (the standard bridge and the bridge in question).

The number of nodes added equals the number of columns added, shown by the red box in panel (a). The number of beams added equals the number of 1’s in the upper diagonal of the right matrix minus the number of 1’s in the upper diagonal of the left matrix (panel (b)). The number of beams moved equals the number of differences (1 changed to 0, or 0 changed to 1), only considering the upper diagonal of the left matrix (columns 1-3 and rows 1-3 of the right matrix in panel (c)). The total edit distance between the two shapes shown in Figure 19 is therefore 1.
Figure 19. Calculating graph edit distance using adjacency matrices

The total edit distance is the sum of the number of nodes added, the number of beams added, and the number of beams moved.
Panel (a): Number of nodes added equals the number of columns added (= 1). Panel (b): Number of beams added equals the number of 1’s in the upper diagonal of the right matrix minus the number of 1’s in the upper diagonal of the left matrix (= 1).
Panel (c): Number of beams moved = Number of differences (1 changed to 0, or 0 changed to 1) in the original upper diagonal (= 1).

Challenges with measuring design “tweaks”

Some design changes affect the bridge’s appearance, but involve the change in the length of a beam rather than the 3 operations listed in Figure 4. For example, moving one node in the standard bridge (and its connected beams) to a different position would affect the length of its connected beams and give the bridge a different appearance, but its graph would be isomorphic to that of the standard bridge (as illustrated in Figure 20).
Measuring these changes in appearance and combining them with structural changes into a
single measure is not straightforward. One issue is how a “tweak” should be counted. For example, according to the principals of graph edit distance, a “tweak” would involve replacing one beam with a beam of a different length. Counted this way, the number of changes for the bridge in the right panel of Figure 20 would be 5. Alternatively, that bridge could be counted as 1 step away from the standard bridge if we only count the number of nodes moved. Another issue is how visually dissimilar a “tweak” must be to be counted and how to measure dissimilarity - the repositioning of the node in the right panel of Figure 20 is clearly visible, but it is unclear what cutoff point we should use to consider a repositioning to be dissimilar.

**Figure 20.** Example of a design “tweak”

In the right panel, the centremost node (blue) has been pulled down so the bridge is visually non-identical to the standard bridge, though the relationships between the nodes and edges are identical to that of the standard bridge.

**Appendix C: Supplementary tables and figures**

**Figure 21.** Bridge building simulation interface
Table 4. Regressions with dependent variable Δcreativity (all designs submitted)

<table>
<thead>
<tr>
<th></th>
<th>Pooled</th>
<th>Control</th>
<th>C-H</th>
<th>C-L</th>
<th>E-H</th>
<th>E-L</th>
<th>ET-H</th>
<th>ET-L</th>
<th>Wald</th>
</tr>
</thead>
<tbody>
<tr>
<td>Stable$_{t-1}$</td>
<td>-0.584</td>
<td>-0.633</td>
<td>-3.621</td>
<td>-0.156</td>
<td>0.193</td>
<td>0.089</td>
<td>-1.201</td>
<td>-0.613</td>
<td>4.72</td>
</tr>
<tr>
<td></td>
<td>(0.357)</td>
<td>(0.587)</td>
<td>(2.266)</td>
<td>(1.181)</td>
<td>(0.556)</td>
<td>(0.814)</td>
<td>(1.031)</td>
<td>(1.224)</td>
<td>(0.581)</td>
</tr>
<tr>
<td>Remaining budget$_{t-1}$</td>
<td>0.006</td>
<td>0.01</td>
<td>-0.016</td>
<td>-0.026</td>
<td>0.015</td>
<td>0.014</td>
<td>-0.001</td>
<td>0.006</td>
<td>16.7</td>
</tr>
<tr>
<td></td>
<td>(0.003)</td>
<td>(0.003)</td>
<td>(0.013)</td>
<td>(0.025)</td>
<td>(0.004)</td>
<td>(0.010)</td>
<td>(0.003)</td>
<td>(0.024)</td>
<td>(0.010)</td>
</tr>
<tr>
<td>Creativity$_{t-1}$</td>
<td>-0.326</td>
<td>-0.304</td>
<td>-0.301</td>
<td>-0.366</td>
<td>-0.344</td>
<td>-0.263</td>
<td>-0.43</td>
<td>-0.43</td>
<td>3.52</td>
</tr>
<tr>
<td></td>
<td>(0.032)</td>
<td>(0.044)</td>
<td>(0.130)</td>
<td>(0.114)</td>
<td>(0.075)</td>
<td>(0.085)</td>
<td>(0.088)</td>
<td>(0.112)</td>
<td>(0.742)</td>
</tr>
<tr>
<td>Design number (t)</td>
<td>-0.014</td>
<td>-0.029</td>
<td>0.002</td>
<td>-0.086</td>
<td>-0.031</td>
<td>-0.007</td>
<td>0.085</td>
<td>0.097</td>
<td>13.08</td>
</tr>
<tr>
<td></td>
<td>(0.017)</td>
<td>(0.029)</td>
<td>(0.069)</td>
<td>(0.087)</td>
<td>(0.020)</td>
<td>(0.038)</td>
<td>(0.046)</td>
<td>(0.043)</td>
<td>(0.042)</td>
</tr>
<tr>
<td>Stable$_{t-2}$</td>
<td>-0.577</td>
<td>-1.595</td>
<td>2.116</td>
<td>3.409</td>
<td>-0.155</td>
<td>-0.536</td>
<td>0.876</td>
<td>-0.256</td>
<td>12.07</td>
</tr>
<tr>
<td></td>
<td>(0.338)</td>
<td>(0.479)</td>
<td>(2.569)</td>
<td>(2.075)</td>
<td>(0.770)</td>
<td>(0.517)</td>
<td>(1.142)</td>
<td>(1.323)</td>
<td>(0.060)</td>
</tr>
<tr>
<td>Remaining budget$_{t-2}$</td>
<td>-0.001</td>
<td>-0.004</td>
<td>0.018</td>
<td>0.025</td>
<td>-0.014</td>
<td>0.005</td>
<td>0.004</td>
<td>-0.026</td>
<td>14.01</td>
</tr>
<tr>
<td></td>
<td>(0.003)</td>
<td>(0.004)</td>
<td>(0.009)</td>
<td>(0.019)</td>
<td>(0.011)</td>
<td>(0.006)</td>
<td>(0.004)</td>
<td>(0.032)</td>
<td>(0.030)</td>
</tr>
<tr>
<td>Creativity$_{t-2}$</td>
<td>0.128</td>
<td>0.118</td>
<td>0.235</td>
<td>-0.03</td>
<td>0.109</td>
<td>0.114</td>
<td>0.261</td>
<td>0.057</td>
<td>8.12</td>
</tr>
<tr>
<td></td>
<td>(0.035)</td>
<td>(0.041)</td>
<td>(0.134)</td>
<td>(0.064)</td>
<td>(0.146)</td>
<td>(0.086)</td>
<td>(0.107)</td>
<td>(0.098)</td>
<td>(0.230)</td>
</tr>
<tr>
<td>Stable$_{t-3}$</td>
<td>0.4</td>
<td>0.715</td>
<td>1.566</td>
<td>-0.076</td>
<td>-1.038</td>
<td>-1.027</td>
<td>0.071</td>
<td>3.258</td>
<td>16.04</td>
</tr>
<tr>
<td></td>
<td>(0.312)</td>
<td>(0.460)</td>
<td>(1.639)</td>
<td>(1.305)</td>
<td>(0.662)</td>
<td>(0.597)</td>
<td>(0.934)</td>
<td>(1.335)</td>
<td>(0.014)</td>
</tr>
<tr>
<td>Remaining budget$_{t-3}$</td>
<td>-0.002</td>
<td>-0.005</td>
<td>0.022</td>
<td>-0.004</td>
<td>-0.007</td>
<td>-0.007</td>
<td>0.002</td>
<td>0.038</td>
<td>21.19</td>
</tr>
<tr>
<td></td>
<td>(0.002)</td>
<td>(0.002)</td>
<td>(0.008)</td>
<td>(0.014)</td>
<td>(0.010)</td>
<td>(0.007)</td>
<td>(0.003)</td>
<td>(0.017)</td>
<td>(0.002)</td>
</tr>
<tr>
<td>Creativity$_{t-3}$</td>
<td>0.024</td>
<td>-0.013</td>
<td>0.04</td>
<td>0.176</td>
<td>0.214</td>
<td>-0.052</td>
<td>-0.051</td>
<td>0.122</td>
<td>16.61</td>
</tr>
<tr>
<td></td>
<td>(0.030)</td>
<td>(0.033)</td>
<td>(0.160)</td>
<td>(0.080)</td>
<td>(0.092)</td>
<td>(0.037)</td>
<td>(0.069)</td>
<td>(0.079)</td>
<td>(0.011)</td>
</tr>
</tbody>
</table>

$R^2$ 0.137 0.156 0.110 0.220 0.183 0.139 0.194 0.231

$N$ 3590 1714 233 186 384 389 431 253

All regressions have change in creativity (Δ) as the dependent variable and include a constant term. Robust standard errors, clustered at the individual level, are reported in brackets. The regression equation was estimated on the full sample of designs (Pooled) and treatment-specific subsamples (C-H = creativity (high), C-L = creativity (low), E-H = cost efficiency (high), E-L = cost efficiency (low), ET-H = extra time (high), ET-L = extra time (low)). The Wald column shows the F-statistic from a Wald test under the null hypothesis that the coefficient of the row variable is the same across all seven subsamples (control and 6 treatment conditions), with the p-value in brackets.
Table 5. Regressions with dependent variable Δcost ($000s, all designs submitted)

<table>
<thead>
<tr>
<th></th>
<th>Pooled</th>
<th>Control</th>
<th>C-H</th>
<th>C-L</th>
<th>E-H</th>
<th>E-L</th>
<th>ET-H</th>
<th>ET-L</th>
<th>Wald</th>
</tr>
</thead>
<tbody>
<tr>
<td>Remaining budget_{t-1}</td>
<td>0.502</td>
<td>0.469</td>
<td>0.704</td>
<td>0.449</td>
<td>0.873</td>
<td>0.444</td>
<td>0.407</td>
<td>0.472</td>
<td>11.97</td>
</tr>
<tr>
<td></td>
<td>(0.077)</td>
<td>(0.090)</td>
<td>(0.099)</td>
<td>(0.151)</td>
<td>(0.153)</td>
<td>(0.065)</td>
<td>(0.152)</td>
<td>(0.152)</td>
<td>0.063</td>
</tr>
<tr>
<td>Creativity_{t-1}</td>
<td>0.139</td>
<td>0.038</td>
<td>1.06</td>
<td>-0.195</td>
<td>1.442</td>
<td>0.07</td>
<td>0.105</td>
<td>-0.718</td>
<td>6.54</td>
</tr>
<tr>
<td></td>
<td>(0.299)</td>
<td>(0.309)</td>
<td>(0.659)</td>
<td>(0.445)</td>
<td>(0.888)</td>
<td>(0.417)</td>
<td>(0.462)</td>
<td>(0.746)</td>
<td>0.366</td>
</tr>
<tr>
<td>Design number (t)</td>
<td>-0.087</td>
<td>-0.279</td>
<td>-0.216</td>
<td>-1.306</td>
<td>-0.121</td>
<td>0.251</td>
<td>0.389</td>
<td>0.86</td>
<td>21.66</td>
</tr>
<tr>
<td></td>
<td>(0.146)</td>
<td>(0.077)</td>
<td>(0.090)</td>
<td>(0.151)</td>
<td>(0.153)</td>
<td>(0.065)</td>
<td>(0.152)</td>
<td>(0.152)</td>
<td>0.063</td>
</tr>
<tr>
<td>Stable_{t-2}</td>
<td>1.744</td>
<td>3.105</td>
<td>2.474</td>
<td>0.413</td>
<td>-2.084</td>
<td>-4.778</td>
<td>11.64</td>
<td>12.028</td>
<td>4.13</td>
</tr>
<tr>
<td></td>
<td>(3.308)</td>
<td>(4.036)</td>
<td>(13.293)</td>
<td>(4.530)</td>
<td>(5.678)</td>
<td>(8.008)</td>
<td>(18.893)</td>
<td></td>
<td>0.659</td>
</tr>
<tr>
<td>Remaining budget_{t-2}</td>
<td>-0.033</td>
<td>-0.001</td>
<td>-0.053</td>
<td>0.171</td>
<td>-0.641</td>
<td>-0.113</td>
<td>-0.039</td>
<td>0.085</td>
<td>9.24</td>
</tr>
<tr>
<td></td>
<td>(0.104)</td>
<td>(0.124)</td>
<td>(0.064)</td>
<td>(0.163)</td>
<td>(0.272)</td>
<td>(0.042)</td>
<td>(0.191)</td>
<td>(0.202)</td>
<td>0.161</td>
</tr>
<tr>
<td>Creativity_{t-2}</td>
<td>0.507</td>
<td>0.493</td>
<td>-0.007</td>
<td>0.255</td>
<td>-0.843</td>
<td>-0.032</td>
<td>0.693</td>
<td>1.126</td>
<td>4.13</td>
</tr>
<tr>
<td></td>
<td>(0.433)</td>
<td>(0.690)</td>
<td>(0.877)</td>
<td>(0.463)</td>
<td>(0.807)</td>
<td>(0.417)</td>
<td>(0.874)</td>
<td>(0.874)</td>
<td>0.659</td>
</tr>
<tr>
<td>Stable_{t-3}</td>
<td>-1.689</td>
<td>-0.095</td>
<td>7.531</td>
<td>-2.824</td>
<td>-2.26</td>
<td>-4.038</td>
<td>6.912</td>
<td>-11.03</td>
<td>7.00</td>
</tr>
<tr>
<td>Remaining budget_{t-3}</td>
<td>-0.127</td>
<td>-0.145</td>
<td>-0.064</td>
<td>-0.197</td>
<td>-0.049</td>
<td>-0.144</td>
<td>-0.035</td>
<td>-0.033</td>
<td>5.58</td>
</tr>
<tr>
<td></td>
<td>(0.024)</td>
<td>(0.032)</td>
<td>(0.071)</td>
<td>(0.138)</td>
<td>(0.071)</td>
<td>(0.104)</td>
<td>(0.061)</td>
<td>(0.472)</td>
<td></td>
</tr>
<tr>
<td>Creativity_{t-3}</td>
<td>-0.429</td>
<td>-0.242</td>
<td>-0.484</td>
<td>0.024</td>
<td>0.07</td>
<td>-0.566</td>
<td>-0.688</td>
<td>-1.092</td>
<td>5.51</td>
</tr>
<tr>
<td></td>
<td>(0.209)</td>
<td>(0.387)</td>
<td>(0.900)</td>
<td>(0.449)</td>
<td>(0.622)</td>
<td>(0.442)</td>
<td>(0.360)</td>
<td>(0.402)</td>
<td>0.480</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>R²</th>
<th>0.273</th>
<th>0.291</th>
<th>0.288</th>
<th>0.296</th>
<th>0.415</th>
<th>0.200</th>
<th>0.324</th>
<th>0.284</th>
</tr>
</thead>
<tbody>
<tr>
<td>N</td>
<td>3590</td>
<td>1714</td>
<td>233</td>
<td>186</td>
<td>384</td>
<td>389</td>
<td>253</td>
<td>431</td>
</tr>
</tbody>
</table>

All regressions have change in cost (Δt) measured in $000s as the dependent variable and include a constant term. Robust standard errors, clustered at the individual level, are reported in brackets. The regression equation was estimated on the full sample of designs (Pooled) and treatment-specific subsamples (C-H = creativity (high), C-L = creativity (low), E-H = cost efficiency (high), E-L = cost efficiency (low), ET-H = extra time (high), ET-L = extra time (low)). The Wald column shows the F-statistic from a Wald test under the null hypothesis that the coefficient of the row variable is the same across all seven subsamples (control and 6 treatment conditions), with the p-value in brackets.
Table 6. Summary statistics of independent variables: Demographics, personality traits, output produced

<table>
<thead>
<tr>
<th>Variable</th>
<th>Mean</th>
<th>SD</th>
<th>Min</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td>Female</td>
<td>0.37</td>
<td>0.48</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Age (5-year bands)</td>
<td>2.13</td>
<td>1.51</td>
<td>1</td>
<td>8</td>
</tr>
<tr>
<td>Income (personal annual before-tax, $000s)</td>
<td>18.51</td>
<td>18.53</td>
<td>0</td>
<td>95</td>
</tr>
<tr>
<td>Left-handed</td>
<td>0.11</td>
<td>0.32</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Exposure to similar activities (sum)</td>
<td>5.60</td>
<td>2.04</td>
<td>0</td>
<td>9</td>
</tr>
<tr>
<td>Exposure to relevant subjects (weighted sum)</td>
<td>2.52</td>
<td>1.71</td>
<td>0</td>
<td>6</td>
</tr>
<tr>
<td>Openness (Big-5)</td>
<td>4.88</td>
<td>1.22</td>
<td>1</td>
<td>7</td>
</tr>
<tr>
<td>Conscientiousness (Big-5)</td>
<td>4.68</td>
<td>1.08</td>
<td>1</td>
<td>7</td>
</tr>
<tr>
<td>Extraversion (Big-5)</td>
<td>3.84</td>
<td>1.36</td>
<td>1</td>
<td>7</td>
</tr>
<tr>
<td>Agreeableness (Big-5)</td>
<td>4.79</td>
<td>1.11</td>
<td>1</td>
<td>7</td>
</tr>
<tr>
<td>Neuroticism (Big-5)</td>
<td>4.19</td>
<td>1.40</td>
<td>1</td>
<td>7</td>
</tr>
<tr>
<td>Grit score (1-5 scale)</td>
<td>3.10</td>
<td>0.68</td>
<td>1.13</td>
<td>5</td>
</tr>
<tr>
<td>Experimental problem-solving approach</td>
<td>4.08</td>
<td>0.83</td>
<td>1</td>
<td>6.42</td>
</tr>
<tr>
<td>Willingness to take risks (0-10 scale)</td>
<td>5.54</td>
<td>2.25</td>
<td>0</td>
<td>10</td>
</tr>
<tr>
<td>Task enjoyment (1-5 scale)</td>
<td>3.72</td>
<td>1.13</td>
<td>1</td>
<td>5</td>
</tr>
<tr>
<td>Number of designs submitted (per individual and task)</td>
<td>5.40</td>
<td>7.08</td>
<td>1</td>
<td>45</td>
</tr>
<tr>
<td>Structurally stable designs (proportion, per individual and task)</td>
<td>0.51</td>
<td>0.42</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Creativity (most creative bridge (edit distance), per individual and task)</td>
<td>19.36</td>
<td>12.46</td>
<td>0</td>
<td>101</td>
</tr>
<tr>
<td>Cost (cheapest bridge ($000s), per individual and task)</td>
<td>347.2</td>
<td>40.3</td>
<td>96.3</td>
<td>400.0</td>
</tr>
</tbody>
</table>

Summary statistics for experiment participants (N=550). “Big-5” variables are personality traits measured using the 15-item Big-5 questionnaire of Lang et al. (2011), based on the five-factor (“Big-5”) personality framework (McCrae and Costa 1987, McCrae and John 1992). Grit score indicates an individual’s tendency to persist until a task is complete, measured using the 8-item questionnaire developed by Duckworth (2007). Experimental problem-solving approach measures an individual’s tendency to approach creative problems via trial and error rather than systematic planning (Nielsen et al. 2008). Willingness to take risks is measured using the 0-10 scale proposed by Dohmen et al. (2011). Gender is an indicator variable that equals 1 if female. Age is measured in 5-year bands, where 1 equals 25 and under, 2 equals 26-30, and so on until 8 (56 or older). Income refers to personal annual before-tax income, excluding any benefits received from the government or receipts from other sources. Left handed is an indicator that equals 1 if the participant normally uses their left hand for writing, where ambidextrous individuals (0.65% of participants) are counted as left handed. Number of relevant subjects is a weighted sum, where subjects that were studied for more than one secondary school year/one university term were given double the weight compared to subjects that were studied for one secondary school year/one university term or less. Relevant subject categories were physical sciences, engineering/computer science, and design. Structurally stable designs are bridges that can bear the weight of a standard-size truck.
Table 7. ANOVA tests for demographic balance across treatments

<table>
<thead>
<tr>
<th>Variable</th>
<th>F-statistic</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Female</td>
<td>1.06</td>
<td>0.381</td>
</tr>
<tr>
<td>Age (5-year bands)</td>
<td>0.61</td>
<td>0.691</td>
</tr>
<tr>
<td>Income (personal annual before-tax, $000s)</td>
<td>1.26</td>
<td>0.279</td>
</tr>
<tr>
<td>Left-handed</td>
<td>0.83</td>
<td>0.532</td>
</tr>
<tr>
<td>Exposure to similar activities (sum)</td>
<td>0.63</td>
<td>0.680</td>
</tr>
<tr>
<td>Exposure to relevant subjects (weighted sum)</td>
<td>0.327</td>
<td>0.867</td>
</tr>
<tr>
<td>Openness (Big-5)</td>
<td>0.18</td>
<td>0.971</td>
</tr>
<tr>
<td>Conscientiousness (Big-5)</td>
<td>0.73</td>
<td>0.600</td>
</tr>
<tr>
<td>Extraversion (Big-5)</td>
<td>0.29</td>
<td>0.921</td>
</tr>
<tr>
<td>Agreeableness (Big-5)</td>
<td>0.51</td>
<td>0.769</td>
</tr>
<tr>
<td>Neuroticism (Big-5)</td>
<td>0.48</td>
<td>0.788</td>
</tr>
<tr>
<td>Grit score (1-5 scale)</td>
<td>0.18</td>
<td>0.971</td>
</tr>
<tr>
<td>Experimental problem-solving approach</td>
<td>0.42</td>
<td>0.837</td>
</tr>
<tr>
<td>Willingness to take risks (0-10 scale)</td>
<td>0.74</td>
<td>0.592</td>
</tr>
<tr>
<td>Task enjoyment (1-5 scale)</td>
<td>3.18</td>
<td>0.008</td>
</tr>
</tbody>
</table>

This table shows that balance was achieved across all treatments and individual characteristics. One exception is task enjoyment, but this result is entirely driven by a higher mean in one treatment (efficiency (high)); removing this treatment from the ANOVA test results in an F-statistic of 1.76 (p = 0.136).

Figure 22. Joint and marginal distributions of creativity and cost (all designs)

The top left and top right panel show the marginal densities of creativity and cost ($000s) respectively; the bottom left panel shows the joint distribution. Creativity is measured using edit distance, defined as the minimum number of operations required to transform one bridge into the pre-specified reference bridge. Orange represents designs that satisfy both essential criteria (cost $400000 or less and be structurally stable), green represents designs that cost more than $400000, blue represents designs that are structurally unstable, and grey represents designs that failed both essential criteria.
**Figure 23.** Creativity and cost efficiency of the most creative bridge design (within-person difference, treatment minus control)

This figure shows the distribution of within-person differences (treatment minus control) in the variable of interest: creativity (left panel) and cost efficiency (right panel) of the most creative bridge design. Creativity is measured as edit distance, defined as the minimal number of operations required to transform one bridge design into a pre-specified reference bridge design. Cost efficiency ($000$s) is measured as the amount of budget remaining ($400000$ minus the cost of production). The boxes indicate the first and third quartiles, the solid line inside each box represents the median, and the “whiskers” represent the range.

**Figure 24.** Creativity and cost efficiency of the cheapest bridge design (within-person difference, treatment minus control)

This figure shows the distribution of within-person differences (treatment minus control) in the variable of interest: creativity (left panel) and cost efficiency (right panel) of the cheapest bridge design. Creativity is measured as edit distance, defined as the minimal number of operations required to transform one bridge design into a pre-specified reference bridge design. Cost efficiency ($000$s) is measured as the amount of budget remaining ($400000$ minus the cost of production). The boxes indicate the first and third quartiles, the solid line inside each box represents the median, and the “whiskers” represent the range.
**Figure 25.** Characteristics of the creative process: Relative weight on creativity and willingness to consider inferior solutions (within-person difference, treatment minus control)

This figure shows the distribution of within-person differences (treatment minus control) in the variable of interest. The left panel shows the relative weight on creativity, which ranges from 0 (solely focusing on maximising cost efficiency) to 1 (solely focusing on maximising creativity). The right panel shows the willingness to consider inferior solutions, which is strictly positive and has no theoretical maximum. Both of these variables were estimated via maximum likelihood. The boxes indicate the first and third quartiles, the solid line inside each box represents the median, the “whiskers” represent the range, and dots represent values that are more than 1.5 standard deviations smaller or larger than the first or third quartile, respectively.
Figure 26. Regression results: Willingness to consider inferior solutions

Panel (a) shows the change in the dependent variable associated with a one-standard-deviation increase in the independent variable. Panel (b) shows the coefficients from a probit regression, where the dependent variable takes the value 1 if individual-level data for both treatment and control conditions are nonmissing (the participant submitted at least three designs). All point estimates are shown with 95% confidence intervals, based on robust standard errors. All regressions are pooled across treatments and include treatment-specific indicator variables (with extra-time-low as the base category), as well as a constant term.
References


