Essays on Knowledge Worker Productivity

by

Samer Charbaji

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Business Administration)
in the University of Michigan
2023

Doctoral Committee:

Professor Roman Kapuscinski, Co-Chair
Professor Stephen Leider, Co-Chair
Assistant Professor Jessica Fong
Professor Tanya Rosenblat

Samer Charbaji
charbaji@umich.edu
ORCID iD: 0009-0007-5106-6279

DEDICATION

This dissertation is dedicated to my wonderful wife whose support and positivity helped me pursue research problems that interest me, to my parents who gave me countless opportunities to learn and grow, and to my siblings and friends who supported me throughout my PhD journey. Obtaining my PhD would not have been possible without all the amazing people in my life.

ACKNOWLEDGEMENTS

I would like to thank my advisors, Steve Leider and Roman Kapuscinski, whose help, time, and effort made this research possible. Their patience and thoughtful feedback helped me pursue a line of research that I find personally fulfilling and for that I am eternally grateful. I want to also thank my committee members Tanya Rosenblat and Jessica Fong for their feedback and for their contribution to my dissertation. Finally, I want to thank Joline Uichanco and Mohamed Mostagir for all their help throughout my PhD journey.

TABLE OF CONTENTS

LIST OF FIGURES

FIGURE

LIST OF TABLES

TABLE

LIST OF APPENDICES

ABSTRACT

Knowledge workers operate in relatively complex settings, where they must often use their judgment and learning capabilities to complete knowledge-intensive tasks. This makes them less suited for the productivity analysis techniques traditionally used in operations management. In my dissertation, I use controlled lab experiments to explore various ways of improving knowledge worker productivity. In the first part of the dissertation, I study how the design features of an enterprise social media platform, a popular form of communication technology that knowledge workers can use to seek help, affects a knowledge worker's helping behavior and productivity. I also elicit the behavioral mechanisms driving effective design features. I find that features that draw on behavioral mechanisms in the form of descriptive social helping norms and reciprocity result in minimal, and in some cases detrimental, effects on help with no effect on performance. In contrast, I find that design features that use goal setting motivation and symbolic rewards to leverage an employee's intrinsic motivation to "be more helpful" can be a surprisingly effective way to promote helping behavior and improve performance. In the second part of my dissertation, I study how varying constraints on the usefulness of a knowledge worker's creative output affects their performance on an originality-focused creative task. I show that low usefulness constraints can be an effective way to improve employee performance by encouraging employees to factor in more usefulness in their creative output. In contrast, moderate and high constraints can result in poor employee performance by respectively causing individuals to create output that is creatively poor or output that is not sufficiently useful. Interestingly, I show that, in such cases, employee performance can be improved by "artificially" lowering the usefulness constraint set or by changing the task goal to emphasize usefulness. In the third part of my dissertation, I study how varying a knowledge worker's ability and freedom to switch between a creative and a repetitive task affects their overall performance. I show that forcing employees to switch tasks can improve creative task performance, but can also lower repetitive task performance, resulting in similar overall performance. Interestingly, I find that giving employees the discretion to switch can result in infrequent task switching, lowering overall performance by lowering their repetitive task performance without increasing their creative task performance. I show that, in such cases, behavioral nudges can be a surprisingly effective way to

improve employee performance by encouraging them to voluntarily switch tasks more often.

CHAPTER 1

Introduction

In the early twentieth century, Fredrick Winslow Taylor's Principles of Scientific Management and Henry Ford's use of assembly lines and specialized tasks challenged the prevailing idea that a "workman can best regulate his own way of doing work." Their principles substantially improved worker productivity and helped revolutionize the way business and manufacturing operations were structured in the United States and the rest of the world.

Since then, the operations management literature has established principles to improve worker productivity in settings typically associated with standardized work, where tasks are often physical and repetitive in nature. In the last few decades, the US economy has steadily shifted more toward service and professional jobs, which are typically more creative and knowledge-intensive and give employees greater discretion in carrying out tasks based on their judgment and learning capabilities. This often makes them more susceptible to an employee's behavioral decision-making biases and less suited for traditional analysis techniques. This has led to calls for research on what is often referred to as "knowledge worker" productivity. In my dissertation, I use experimental methods to study behavioral processes that can affect a knowledge worker's productivity in a variety of settings. I conduct my experiments in a controlled lab environment that allows me to establish a causal relationship between a specific behavioral intervention and improvements in productivity. My dissertation consists of three separate research projects which are presented in Chapter 2, Chapter 3, and Chapter 4. I describe each research project briefly below.

In my second dissertation chapter, I explore how different design features of enterprise social media platforms (ESMPs) affect a knowledge worker's helping behavior and productivity and I elicit the behavioral mechanisms driving effective design features. To study this, I give participants a real-effort task to do and a platform that they can use to their request and give help to other participants in real-time. I, then, vary the design features of the platform across treatments and study the effect that has on a participant's helping behavior and productivity. The treatments I run are inspired by key design features found on popular real-world ESMPs. I run two treatments that vary the visibility of helping information on the

platform and one treatment that sets helping goals that participants can pursue and private symbolic badges they can obtain. My results show that both information treatments, which leverage a form of descriptive social norms and a form of reciprocity, result in minimal, and in some cases detrimental, effects on help with no effect on performance. I find that helping goals and badges are surprisingly effective at improving a participant's helping behavior, by leveraging their intrinsic motivation to help.

In my third dissertation chapter, I study how increasing usefulness constraints affect the performance of knowledge workers on an originality-focused creative task and if managers can benefit from "artificially" varying the usefulness constraint they set for their employees. I explore this in a lab setting where participants are assigned a creative task to work on and are incentivized based on the originality of their output on condition that they meet a usefulness constraint that I vary across treatments. As expected, I show that setting a low usefulness constraint is an effective way improve participant performance subject to that constraint by encouraging participants to factor in more usefulness in their creative output. Surprisingly, I find that setting a moderate usefulness constraint causes participants to produce output that is creatively poor, resulting in poor participant performance. My results show that participant performance, subject to the moderate constraint, can be improved by artificially lowering the constraint set to participants. I also find that setting a high usefulness constraint is not effective at encouraging participants to sufficiently emphasize usefulness, which also results in poor performance. My results show that participant performance, subject to the high constraint, can be improved by artificially changing the goal to emphasize usefulness rather than originality.

In my fourth dissertation chapter, I explore how varying a knowledge worker's ability and freedom to switch between a creative task and a repetitive task affects their performance on each task and their overall performance. I study this in a lab setting by giving participants a creative task and a repetitive task to work on and by either forcing them to switch tasks repeatedly, forcing them not to switch tasks, or by giving them the discretion to switch tasks. My results show that forcing participants to switch tasks improves their performance on the creative task by helping them get "unstuck" finding a solution to a creative question. Doing so does reduce their performance on the repetitive task, which results in similar overall performance to forcing them not to switch. Interestingly, I find that participants, given the discretion to switch, rarely switch tasks. This lowers their overall performance by lowering repetitive task performance without increasing their creative task performance. I show that participant performance, in such cases, can be improved by setting behavioral nudges that encourage participants to voluntarily switch between tasks more often.

CHAPTER 2

Enterprise Social Media Platform Design and Knowledge Worker Productivity

2.1 Introduction

Enterprise social media platforms (ESMPs) are company-hosted, web-based platforms that make it easier for employees to communicate. ESMP adoption is widespread with 85% of respondents in a recent McKinsey Global Survey indicating that their companies use such technologies (Bughin et al., 2017). One of the key reasons for implementing ESMPs is to encourage help and collaboration between employees, but their implementations suggest mixed success, with a recent study showing that less than half of implementations have many employees that use them regularly (Charki et al., 2018; Li, 2015). Despite mixed reviews, many companies remain interested in implementing ESMPs to promote innovation and to compete with peer firms (Bughin, 2015). Prominent platforms, such as Facebook Workplace and Jive, typically mimic the designs of public social media platforms, such as Facebook, to make them more familiar for employees to use. Importantly, these platforms often differ in some of the design features they add to promote employee usage, however, existing research provides little guidance on which individual design features are most effective (Van Osch et al., 2015). Our paper aims to help companies implementing ESMPs identify specific platform design features that can effectively improve employee helping behavior and performance and to understand the behavioral mechanisms driving their effectiveness.

The existing literature on ESMPs has focused on defining them and on discussing their design and how they differ from previous enterprise communication technologies, such as blogs and wikis (Leonardi et al., 2013). The literature has also used survey and observational data to understand factors affecting an employee's decision to use these platforms and the patterns of employee helping behavior on these platforms (e.g. Brzozowski et al., 2009; Cardon and Marshall, 2015; Stieglitz et al., 2014; Rode, 2016). Our paper compliments the

ESMP literature by studying the design features of ESMPs in a controlled lab environment, which allows us to establish a causal relationship between a specific platform design feature and changes in employee helping behavior and productivity. Our experiment allows us to then vary the design of effective features to identify the potential behavioral mechanisms driving their effectiveness. The results from our paper also speak to a separate literature that has studied the effect of "gamified" designs (ones that include game-like design elements) on user activity and engagement in real-world public Q&A platforms such as Stack Overflow, a popular platform for software developers (Srba and Bielikova, 2016; Oktay et al., 2010). This literature finds that gamified designs are often an effective way of increasing user activity and engagement (see Grant and Betts, 2013), however, this literature also relies on observational data and interviews, which makes it difficult the elicit the effectiveness of separate design features.

Our experiment intends to mirror work situations where knowledge workers are constrained on time and must balance between working on their, often non-repetitive and knowledge-intensive, tasks and spending their time helping others on the platform (see Gibbs et al., 2013). We replicate this trade-off by constraining participants on time, incentivizing them based on their performance on their assigned non-repetitive and knowledge-intensive tasks, and giving them a platform they can spend their limited amount of time on to request and give help to other participants in real-time. We conduct two experimental studies. In Study 1, we explore the effectiveness of separate ESMP design features on helping behavior and performance and, in Study 2, we vary the design of our most successful design feature from Study 1 to identify the behavioral mechanisms driving its effectiveness.

Study 1 consists of four treatments. The baseline treatment [Ø] provides a basic platform that allows participants to request, receive, and give help to other participants in their group. The next two treatments increase the visibility of helping behavior on the platform. In the aggregate information treatment [AI], each participant can see the total amount of help requested and given by the group and their own helpfulness rank compared to the group. This may encourage the participants to alter their helping behavior in response to the group's helping dynamics. The full information treatment [FI] augments the aggregate information (found in the [AI] treatment) with individual information about the helping behavior of each group member. This may prompt participants to focus their help on certain group members, e.g. helpful ones. The badges treatment [B] takes a different approach to increase motivation to use the platform. This treatment sets two helping goals on the platform and rewards participants with a badge for achieving each goal that is private (only visible to them) and symbolic (does not affect their final payoff). Participants can complete the first goal by giving and requesting help early on in a round and the second goal by giving

enough help by the end of a round. This can encourage participants to increase their helping behavior in pursuit of these helping goals and badges. By comparing participant behavior across treatments, we aim to evaluate the design features that best promote helping behavior and increase productivity.

Our Study 1 treatments are inspired by design features found on ESMPs such as Facebook Workplace, Jive, and Microsoft Yammer. Our [AI] treatment is designed to give participants a general signal of their group's helping behavior. This resembles the effect of adding "trending" posts on ESMPs. While the main function of trending posts is to bring employee attention to particular posts generating significant conversation in the community, doing can also allows employees to see the volume of employee activity associated with these posts. This can help employees form an impression of general employee use of the platform, which can encourage them to use the platform more. Our [FI] treatment is designed to give participants a signal of each group member's helpfulness. This resembles design features found on platforms, such as Jive, that allow employees to accrue points and levels from using the platform, which appear publicly on their profile and that can, in part, signal an employee's helpfulness on the platform to others. Finally, our [B] treatment encourages participants to help by adding helping goals and symbolic badges. This resembles design features found on platforms, such as Facebook Workplace, that reward employees with badges for completing specific goals on the platform. It is important to note that the badges employees earn in real-world platforms often appear publicly on the their profile, which can, similarly to points and levels, signal an employee's helpfulness to others. In our setting, we focus exclusively on the role of goals and badges in leveraging an employee's intrinsic motivation to help others and we do so by making the symbolic badges participants earn private (i.e. not visible to other participants).

The results from Study 1 show that establishing helping goals with private and symbolic rewards in the [B] treatment is the most effective way of increasing both helping behavior and productivity. The [B] treatment nearly doubles help given, compared to the basic platform, and leads to a significant increase in participant performance. Surprisingly, both information treatments have neutral to negative effects on helping behavior with no effect on productivity. Given the dramatically higher effectiveness of the [B] treatment over the "information" treatments, we focus on identifying the behavioral mechanisms driving its effectiveness in Study 2.

The [B] treatment's effectiveness could be because of goal setting motivation, symbolic reward attainment, nudging participants to use the platform earlier, and/or giving participants intermediate goals and feedback. In Study 2, we run treatments that allow us to

separate the effect of these mechanisms. We show that the effectiveness of the [B] treatment is primarily driven by goal setting and symbolic rewards that leverage a participant's intrinsic motivation to be "more helpful" rather than early helpfulness nudging and intermediate goals and feedback that respectively seek to alter a participant's behavior or reward their behavior in a certain way. Our results are surprising since our time-constrained participants accept and pursue both helping goals even though doing so slightly decreases their payoff (by having them spend their limited time helping others) and rewards them with badges that are symbolic (do not affect their payoff) and private (do not affect how they are perceived by other participants).

2.2 Literature Review

Our paper builds on a number of literature streams in operations management, information systems, and behavioral economics. We, first, discuss the operations management literature studying knowledge worker productivity. We, then, describe the literature studying ESMPs, focusing on the sub-literatures studying employee helping behavior and contribution decisions. Finally, we leverage the behavioral literature on intrinsic and prosocial motivation to make predictions about participant behavior in each of our experimental treatments.

2.2.1 Knowledge Worker Productivity in Operations Management

The operations management literature has separately looked at how knowledge worker productivity is affected by technology fit and employee helping behavior. The literature on technology fit suggests that the complexity of IT-worker systems should match the repetitiveness and novelty of workers tasks (Napoleon and Gaimon, 2004; Ponsignon et al., 2011; Bardhan et al., 2007). Our paper contributes to the technology fit literature by studying how ESMPs, which are a form of productivity technology, can be designed to improve knowledge worker productivity.

The literature on helping behavior has studied the factors affecting knowledge worker interactions, as individuals or teams, across the organization (Roels and Su, 2013; Crama et al., 2019; Schlapp et al., 2015; Song et al., 2018). In these settings, reputational concerns and properly designed incentives encourage members to collaborate and learn from other members, thereby improving their productivity. We contribute to the literature by studying how ESMP design features can, possibly leveraging different behavioral mechanisms, promote help among knolwedge workers across the company.

2.2.2 Enterprise Social Media Platforms

We adapt our definition of ESMPs from Leonardi et al. (2013) and define them as platforms that allow employees to make and reply to easily viewable posts and which may also include social features, reputation systems, or rewards and incentives.

The literature on ESMPs has mainly relied on interviews, surveys, and observational data to study employee behavior patterns on the platform and factors affecting an employee's decision to contribute to the platform. Research has shown that factors such as company culture or manager and coworker feedback can affect employee contribution decisions, while considerations such as user activity and similarity of expertise can affect who employees are likely to respond to on the platform (Cardon and Marshall, 2015; Brzozowski et al., 2009; Stieglitz et al., 2014; Hwang et al., 2015; Bulgurcu et al., 2018).

Our paper contributes to the ESMP literature by introducing a novel experimental setting that allows us to establish a causal relationship between an ESMP design feature and changes in helping behavior and, in turn, changes in performance. Our experimental design also allows us to elicit the behavioral mechanisms driving the increases in helping behavior we see in effective design features. This can help us understand why some popular ESMPs choose to include certain design features and explain some of the behavioral mechanisms driving the effectiveness of these features.

2.2.3 Behavioral Literature on Intrinsic and Prosocial Motivation

We draw on insights from the behavioral literature to suggest potential behavioral mechanisms that may support the efficacy of the design features we test. In our review of the literature, we discuss experiments studying public-good games, social norms, reciprocity, goal setting, symbolic rewards, and gamification.

2.2.3.1 Public Goods.

Our setting is closely related to public-goods experiments. Participants in our experiment are constrained on time, paid solely based on the number of tasks they complete, and given a platform, they can spend their limited time to request and give help to other participants in their group. Giving help benefits others, but leaves participants with less time to work on their own tasks (decreasing their payoff), while requesting and receiving help helps them solve tasks more quickly (increasing their payoff). The intended design of our experiment is such that groups that request and give more help will have a higher performance. We, however, leave it up to each participant to decide if and to what extent they want to use the

platform to request and/or give help. Naturally, participants, in our setting, may be tempted to request help without giving help. This mimics the challenge of free-riding behavior common in public-goods games, where a participant benefits without contributing to the public good. This risk of free-riding commonly deters participants from cooperating in public goods games and effectively stops them from reaching the social optimum (see Kagel and Roth, 2016 and Chaudhuri, 2011 for surveys on public goods games). Based on the public-goods literature, we expect participants in the [Ø] treatment, who do use the platform, to free-ride by requesting and not giving sufficient help. We, then, consider other treatments that add design features intended to reduce free-riding behavior by encouraging participants to give more help.

2.2.3.2 Descriptive Social Norms.

Our two information treatments add design features that give participants varying levels of information on the helping behavior of other participants. Specifically, the [AI] treatment gives participants aggregate-level helping information by allowing a participant to see the number of times she requested, received, and gave help, the total amount of help requested and given on the platform by her group, and her helpfulness rank within the group. This feature is designed to help participants infer the group's descriptive social norms on helping - a common understanding of how one should behave (or help) in our experiment (Cialdini, 2007). If participants are motivated to conform to the descriptive norm, access to aggregate-level helping information may limit free-riding behavior by encouraging participants to give more help. Relating this back to public goods experiments more directly, Keser and Van Winden (2000) find that a majority of participants in public goods experiments are conditional cooperators, who increase (decrease) their contribution in one round depending on whether they were below (above) the average contribution in the previous round. For more on social norms in public-goods games, we refer the reader to Fischbacher et al. (2001).

2.2.3.3 Reciprocity.

Reciprocity motivates individuals to sacrifice resources and to be kind to those who are kind, or to punish those who are unkind (Rabin, 1993). Our [FI] treatment gives participants access to individual helping information and the IDs of those who helped them. Participants can use this information to direct their help toward a particular participant as a reward or away from a participant as punishment. In a public goods setting, Fehr and Gächter (2000) show that introducing costly punishment, where a participant pays a fee to significantly decrease another person's payoff, deters uncooperative behavior and moves the group closer

to the social optimum of everyone contributing their entire endowment. If the [FI] treatment effectively increases helping behavior, this could be explained by participants adjusting their helping behavior by targeting their help towards helpful individuals and away from unhelpful individuals, which may help mitigate free-riding and increase overall help. We will also check for generalized reciprocity, where an individual responds to kindness by being kind to someone else, or to the group as a whole (Baker and Bulkley, 2014).

2.2.3.4 Goal Setting and Symbolic Rewards.

Our badges [B] treatment adds design features that leverage insights from both the goal setting and symbolic rewards literatures. Goal setting is often an effective way of shaping employee behavior in organizations (Latham and Locke, 2006). The literature finds that setting difficult, but achievable goals, if accepted, results in higher performance than general or "do your best" goals. For an excellent review of the goal setting literature, we refer the reader to Locke and Latham (2002). The literature argues that for goal setting to be effective, employees should be willing to accept and commit to a set goal, which depends, in part, on their perceived benefit from the goal (Erez and Kanfer, 1983). For example, Erez and Zidon (1984) find that increasing goal difficulty results in higher participant performance, if the goal is accepted, and in lower participant performance, if the goal is rejected.

In Study 1, the badges [B] treatment sets an "early help" and a "quantity of help" goal to encourage participants to give more help. Participants, in our experiment, are paid based solely on the number of word puzzles they solve correctly, regardless of the amount of help they give (or the helping goals they complete) on the platform. Unlike the goals often considered in the literature, pursuing either of the helping goals requires a participant to spend time giving help to other participants, leaving them with less time to work on their own word puzzles and likely decreasing their final payoff. Participants, in this case, may not see a benefit from pursuing either helping goal and may choose to not give help on the platform.

Receiving a badge may encourage participants to pursue the helping goals we set. Research has shown that awarding employees non-monetary symbolic rewards can increase employee effort (Kosfeld and Neckermann, 2011; Gallus, 2017; Bradler et al., 2016). The symbolic rewards often considered in the literature are public (visible by other employees) and rely on reputation building, competition, and/or recognition. Unlike the literature, the symbolic badges we set are private, relying entirely on a participant's intrinsic motivation to help. As such, participants in our experiment may not see the benefit of earning our badges and the presence of private and symbolic badges may not encourage participants to pursue

the helping goals we set.

If the [B] treatment is effective in increasing helping behavior, it could be because of goal setting and/or symbolic rewards that encourage participants to be more helpful, or because of the game-like elements of our goal design, which we discuss in the gamification section below. In Study 2, we run additional treatments to help us disentangle the effect of goal setting, symbolic rewards, and game-like mechanics on helping behavior in the [B] treatment.

2.2.3.5 Gamification.

Gamification refers to the use of game elements in non-game settings (see Deterding et al., 2011), typically to increase motivation towards a desired behavior. Successful gamification often relies on goals and rewards that reinforce desired user behavior to encourage its repetition (Seaborn and Fels, 2015). The gamification literature suggests that properly spaced and structured goals and rewards can increase the effectiveness of gamification by giving users easier goals that they can complete with less effort and more difficult goals they can then pursue after receiving positive feedback and reinforcement (Marder, 2015; Anderson et al., 2013).

The helping goals set in our [B] treatment are inspired by insights from the gamification literature (Bista et al., 2012). Participants in the [B] treatment can, in each round of the social media stage, be rewarded with an "early help" badge for requesting and giving help once in the first two minutes of a round and a "quantity of help" badge for giving help five times in a round. The "early help" badge is designed to encourage participants to start using the platform early in the round by giving them a sub-goal that benefits them (request help once) and a sub-goal that benefits others (give help once). Encouraging participants to use the platform early gives participants more time to request and give help during the rest of the round. Also, completing both sub-goals can help participants establish norms to request and give help on the platform. To disentangle the effect of early nudging and/or early norm establishing from overall helpfulness nudging, we separately run an "early help"-only badge treatment and a "quantity of help"-only badge treatment in Study 2.

Completing the "early help" badge also helps participants make progress towards completing the "quantity of help goal". As such, the "early help" goal and badge may be considered a form of intermediate goal and feedback that encourages participants to pursue the relatively more demanding "quantity of help" goal. To test the effect of intermediate goals and feedback and their spacing more directly, in Study 2, we run an additional badges treatment that adds an intermediate "help 3 times" quantity of help goal and badge to the [B] treatment. If intermediate goals and feedback are a major driver of helping behavior in

the [B] treatment, we would expect that adding this intermediate goal would increase the proportion of participants helping beyond the "early help" badge and/or pursuing the "help 5 times" badge, and would increase overall helping behavior.

Our paper contributes to the streams of literature discussed above in several respects. We show that adding an enterprise social media platform with certain design features can be an effective way to promote help among knowledge workers and improve productivity. Our approach differs from previous studies on ESMPs by using a laboratory experiment to separate ESMP design features that are often bundled together in real-world platforms, allowing us to first establish a causal relationship between the presence of a design feature and resulting changes in a participant's helping behavior and then to elicit the underlying behavioral mechanisms driving effective design features. We show that the effect of additional helping information is small, with aggregate helping information moving helping behavior closer to the average and individual helping information having little effect on helping behavior. Importantly, we show that adding helping goals with private and symbolic badges, that leverage a participant's intrinsic motivation to help, is a surprisingly effective way of promoting helping behavior and improving performance. Our results contribute to the literature on goal setting and symbolic rewards by showing the effectiveness of both in promoting helping behavior, even when the helping goals set are not tied to a direct improvement in a participant's own performance (in fact, they leave participants with slightly less time to work on their own tasks) and even when the symbolic rewards awarded are private and do not contribute to a participant's reputation or public image.

2.3 Study 1: Experimental Design

Our experiment is designed to mirror real-world settings where a knowledge worker is working on a non-repetitive knowledge-intensive task that they can most likely, with sufficient time, solve correctly, but which they can often solve more efficiently if they ask for help. For example, an engineer trying to debug a piece of code may be able to do so by directly guessing where the bug is by looking at their code and their compiler's error message or they could, with sufficient time, find the bug by going through their code "line-by-line." Alternatively, they can ask for help my making a post about the code on the company's ESMP that other employees can view and that another employee who has encountered a similar issue before can answer relatively quickly.

To mirror such real-world settings, our experiment asks participants to repeatedly perform a real-effort task solving word puzzles. Participants work across two stages: in the first stage (three rounds) subjects perform the task individually, while in the second stage (five

rounds) a platform is introduced allowing subjects to ask for and receive help from five other participants. Each assigned word puzzle question consists of a factual sentence that is missing one word, which is given to participants with its letters scrambled. A participant's task is to write out a word answer for each question and submit it. The computer will automatically inform a participant if their submitted answer is incorrect, allowing them to try again. A participant can guess the correct word answer by using clues from the factual sentence and/or by using the word's scrambled letters. They can also resort to "brute forcing" an answer by continuously submitting different combinations of the word's letters, which, although possible, is likely to take a significant amount of their time. In Stage 2, participants can also use a platform to either request or give help to other participants in real-time as they work on their own assigned tasks. Using the platform is designed to help participants work on their tasks more efficiently.

At the end of the experiment, participants are paid based on their total performance, which is measured as the number of word puzzle questions they solve correctly. We use a participant's performance in Stage 1 as a measure of their individual ability, which we then control for, when examining participant performance and helping behavior in Stage 2. Having four treatments corresponding to different design features of the platform in Stage 2, allows us to establish a causal relationship between the addition of a design feature and the resulting changes in participant helping behavior and performance. At the end of the experiment, we also conduct a small survey at to collect participants' demographic information and ask them about their approach and their helping strategies.

2.3.1 Participants

We recruited 264 participants for Study 1. The experiment is implemented using zTree and participants are recruited using ORSEE (Fischbacher, 2007). Participants are paid a $5 show-up fee in addition to a payoff based on the number of word problems they solve correctly. Participant payments ranged from $10 to $29.

2.3.2 Platform Design

Participants are assigned random IDs at the start of Stage 2 that represent them on the platform. They can request help by pressing the "Request Hint" button, which creates an indicator for other participants that the participant is asking for help. Participants can only have one unanswered request at a time (to avoid "spamming" the platform), but can continue working on their word puzzles while they wait for help. They can also cancel their request at any time using the "Cancel Hint" button. The help a participant receives

Table 2.1: Treatment Design of Study 1

Feature	Basic Platform Treatment [Ø]	Aggregate Information Treatment [AI]	Full Information Treatment [FI]	Badges Treatment [B]
Ability to Request and Give Help	✓	✓	✓	✓
Helpfulness Rank in Group		✓	✓	
Total Help Requested and Given by Group		✓	✓	
Help Requested, Given, and Received by Each Participant			✓	
ID of Participant that Gave Participant Help			✓	
Helping Goals that Update with Progress				✓
Private Badge for Completing a Helping Goal				✓

Note: Participants, across all treatments, first spend three rounds working individually, allowing us to control for their individual ability when studying the effect of platform features on their performance.

consists of the final word answer minus a few letters, which forces them to work further to complete the problem. This reflects real-world settings where employees typically receive answers to their ESMP posts that would help them rather than complete their work for them. Participants can answer help requests by pressing the "Give Hint To" button on the platform corresponding to the participant requesting help (e.g. "Give Hint To Participant 1"). Requesting and giving help each costs a participant two seconds of their round time (we set an equal cost for simplicity) and are done abstractly to avoid issues with the variability in the opportunity cost of time and the quality of help. The number of tasks, difficulty of tasks, helping cost, hint format, and round time are designed so that a participant is very unlikely to finish all 30 questions on their own (on average they solve 18 out of 30 questions) and can benefit from receiving help. This creates a situation where each group member has to balance between spending time working on their own tasks and time requesting and giving help on the platform. In practice, employees using ESMPs can easily see help requests from different

employees at the organization, but they would typically only be able to answer a subset of questions they are knowledgeable about. As such, participants in our experiment can only answer requests if they see that they are "eligible" to answer them. They are eligible to answer a request if they previously answered that question or a similar one. For each question in the social media stage, we make two participants eligible to answer a request. Fixing this number ensures that participants have an equal probability of getting an answer no matter which question they ask and avoids participants needing to form beliefs about whether a lack of help comes from insufficient desire or insufficient ability.

2.3.3 Procedure

Participants are anonymous and do not know who they are partnered with on the platform. They are not allowed to speak to each other during the experiment and can only interact by requesting and giving help. At the start of the experiment, instructions are both displayed as text and read aloud. Participants are given a set of practice exercises to make sure they knew how the word unscrambling works and how their payoff is going to be calculated. At the start of Stage 2, additional instructions and practice exercises are given, illustrating the use of the social media platform.

2.3.4 Treatments

The four treatments we consider differ in the design features of the social media platform (see Table 2.1). The basic platform [Ø] treatment provides a "bare bones" platform that allows participants to request help and give help. Our two information treatments add features that provide participants with information on other participants' helping behavior. The "Aggregate Information" [AI] treatment shows a participant the total number of help requests made and the total number of help requests answered by the group from the start of the social media stage, and privately shows her the number of times she requested, received, and gave help from the start of the social media stage. Participants are also privately shown their helpfulness rank (visible only to them) calculated based on the difference between the number of times they requested help and the number of times they gave help. The "Full Information" [FI] treatment provides the same information, as well as information about each specific participant; the number of times each specific participant requested, received, and gave help. If a participant receives help, she is also able to see the ID of the participant who helped her next to the hint she received. The platform displays all the hints received and the IDs of the participants who answered the help requests from the start of the social media stage. Helping information and rank in both information treatments is updated every

time any of the six participants presses the request help, cancel request, or one of the give help buttons.

The "Badges" [B] treatment builds on the basic platform [Ø] by rewarding participants that complete helping goals with badges that are *private* (cannot be seen by other participants) and *symbolic* (do not affect her final payoff). At the start of the second stage, all participants automatically earn a badge for completing the first stage and accessing the social media platform, which introduces them to the notion of earning badges. In each round, a participant sees two helping goals whose requirements are updated every time they request and/or give help. Participants can then earn the "early help" badge and a "quantity of help" badge in each of the five rounds. Once a participant earns a badge, she continues to see that badge on the platform in all subsequent rounds. Participants earn the "early help" badge by requesting help once and giving help once in the first two minutes of the round. To avoid gaming the system, participants are informed that making a post and canceling it within 30 seconds does not count towards earning the badge. Participants earn the "quantity of help" badge by giving help five times in a round. We set the threshold for the "quantity of help" badge to be substantially higher than the observed average of 1.5 times per round in the [Ø] treatment.

2.4 Study 1: Experimental Results

Our experimental treatments study the effect of separate platform design features on participant performance and helping behavior. We measure participant performance and participant helping behavior respectively as the number of questions answered correctly and amount of help given per round in the social media stage (Stage 2). Table 2.2 presents basic summary statistics of performance and helping behavior across treatments and, importantly, shows that changing the design of the platform (such as in the badges treatment) can result in improvements in participant helping behavior and performance. It also points to heterogeneity in participants' ability to work on our real-effort task, which we control for in our analysis.

2.4.1 Participant Performance

Our first goal is to compare participant performance in the social media stage across treatments. To do so, we run a panel regression of participant performance on treatment dummies, while accounting for participant random effects and using standard errors clustered at the group level. The results of the regression are shown in Table 2.3. Because our treatment

Table 2.2: Summary Statistics

Treatment	**Average Performance**	**Individual Ability**	**Average Help Requested**	**Average Help Given**
Basic Platform	18.50	16.93	1.54	1.45
Aggregate Information	20.39	18.71	1.01	0.96
Full Information	20.09	18.95	1.79	1.73
Badges	21.21	18.89	3.25	3.17

differences could be driven by differences in individual ability across treatments, we add controls for individual ability in Column 2. We find that, when controlling for individual ability, the [B] treatment results in significantly higher performance, a 5% increase, over the [Ø] treatment, while neither of the [AI] or [FI] information treatments significantly improves performance.

Recall that we set a nominal cost of two seconds for requesting or giving help to make a participant's decision to use the platform non-trivial, but wanted receiving help to both increase the chance of answering that question, and on-net boost overall performance. Although not intended by our experimental design, participants might still find requesting and receiving help on the platform too costly to overall performance, since it involves them deciding if a question is worth asking about, making a request about it and incurring the time cost, switching back to the question when the request is answered and working out the final answer based on the hint, and then resuming their work. These mental switching costs could undermine the productivity benefit of help. We find that indeed, at the question level, receiving help is beneficial, increasing the chance of answering a question correctly by 10% (71% with a hint versus 61% without, ranksum test: $p < 0.01$). In Table 2.3, we show that when we control for help requested (in Column 3) and help received (in Column 4), the indicator for the [B] treatment shrinks towards zero and is no longer significant, suggesting that the [B] treatment works primarily through increasing help. As a way of benchmarking the magnitude of the effect of differences in help, if we multiply the difference in help between treatments (3.17 versus 1.45) by the coefficient on help from either Column 3 or 4 in Table 2.3, the predicted effect of the increased help in [B] is approximately 75% as large as the overall treatment difference of the [B] treatment estimated in Column 2.

Table 2.3: Panel Regression: Number of Questions Answered per Individual across Treatments and Rounds

	No Controls	w/ Indiv. Ability	w/ Indiv. Ability & Help Requested	w/ Indiv. Ability & Help Received
AI Treatment	1.890*	0.320	0.527	0.518
	(1.075)	(0.455)	(0.392)	(0.389)
FI Treatment	1.588	-0.190	-0.281	-0.293
	(1.097)	(0.372)	(0.310)	(0.310)
B Treatment	2.710***	0.984***	0.342	0.319
	(1.035)	(0.374)	(0.340)	(0.345)
Individual Ability		0.880***	0.876***	0.876***
		(0.0281)	(0.0269)	(0.0269)
Help Requested			0.379***	
			(0.0490)	
Help Received				0.389***
				(0.0517)
Round Controls	Yes	Yes	Yes	Yes
Constant	17.97***	3.072***	2.728***	2.736***
	(0.880)	(0.639)	(0.583)	(0.578)
Observations	1,320	1,320	1,320	1,320
Number of Subjects	264	264	264	264

Clustered standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

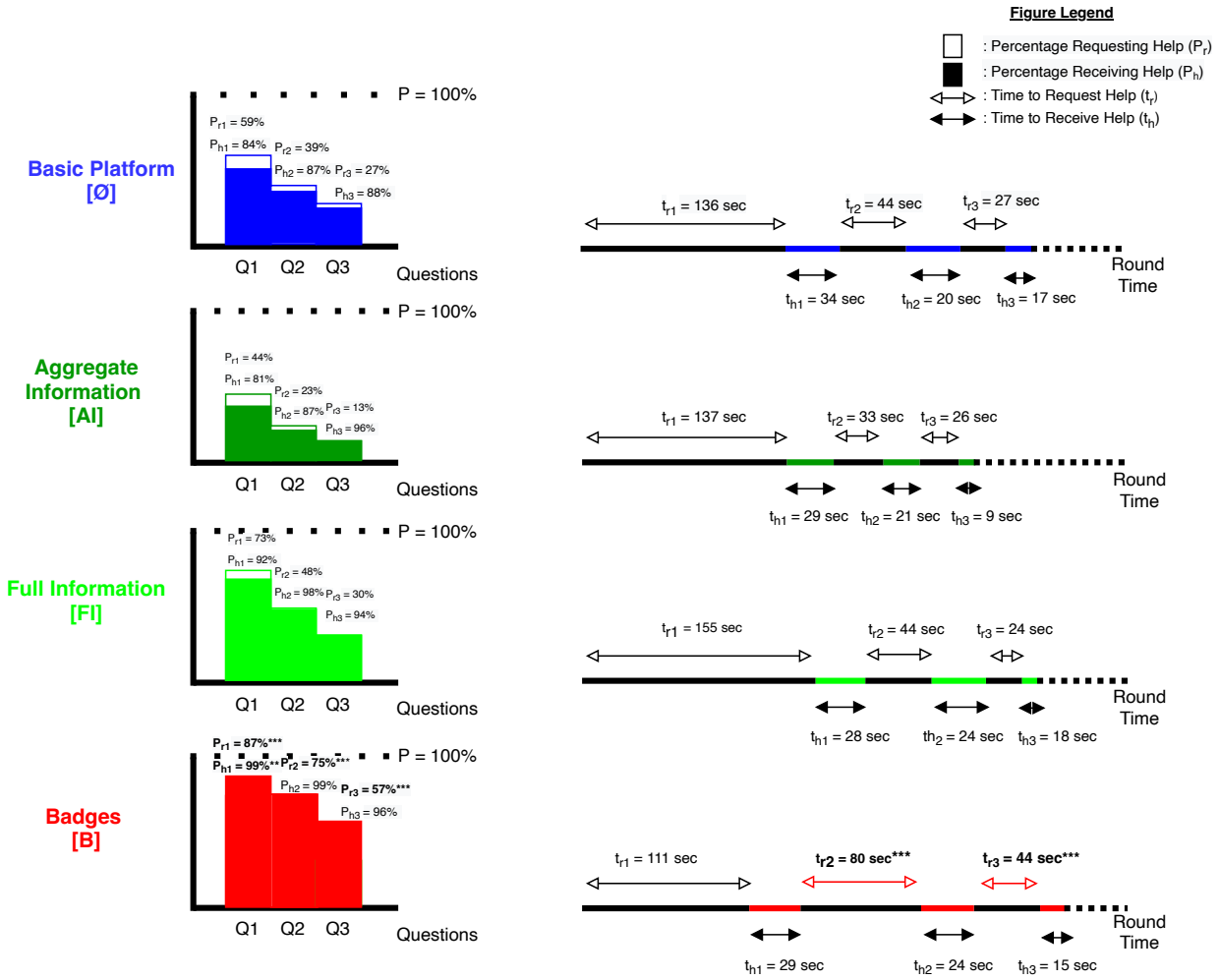
Notes: The coefficients from running a panel regression with random effects and clustering at the group level are reported. The dependent variable is the performance of an individual in a round measured as the number of questions she answered correctly.

2.4.2 Participant Helping Behavior

Our goal now is to understand the differences in the amount of help given and the dynamics of requesting and giving help across treatments. Our results show that the average amount of help given more than doubles in [B] compared to [Ø] treatment (3.17 versus 1.45 posts answered, with rank sum test p=0.00). The difference is not significant in either the [AI] or [FI] treatments compared to [Ø] treatment (rank sum test p=0.17 and p>0.2). We also find similar results when we run regression on participant helping behavior in Section A.1 of the appendix.

To understand the differences in helping behavior, we look at the percentage of partici-

Figure 2.1: Helping Dynamics.



Notes: For each treatment, the percentage of participants that request help (P_r) and the percentage of participants that receive help (P_h) are shown on the left and the average time to request (t_r) and receive help (t_h) are shown on the right for the first three requests (denoted with subscripts 1 to 3). Treatment comparisons are compared to the [Ø] treatment and are shown in bold text if significant : *** p<0.01, ** p<0.05, * p<0.1.

pants requesting help, the time to request help, the percentage receiving help, and the time to receive help for the first three requests across treatments, shown in Figure 2.1. We find no significant difference in any of these measures between the information treatments and the [Ø] treatment. Interestingly, we find a significant increase for the [B] treatment in the percentage of participants that request help (from 59% to 87%) and in the percentage that receive help (from 84% to 99%), but no significant difference in the time to first request help

and time to receive help.[1] Our results show even bigger differences in the probabilities of requesting and receiving help for the second and third time in the round. We deduce that the effectiveness of the [B] treatment comes from increasing the propensity of participants to request and give help, rather than decreasing the time it takes participants to request or give help. We revisit this result in Study 2 where we find no significant effect of "early helpfulness nudging" on the total quantity helping behavior even though it successfully encourages participants to use the platform earlier in the round, thereby giving them "more time" to request and give help in the round.

In the appendix, we examine two additional features of participants' helping behavior: the relationship between help and individual ability in Section A.3, and the effect of receiving help in one round on helping activity in the following round in Section A.4. To briefly summarize these results, we generally find that participants with higher individual ability give more help than participants with lower individual ability, but request help similarly. We also find that receiving help in one round increases the amount of help requested in the following round, but not the amount of help given. This suggests that experiencing the benefit of help encourages participants to request more help, but does not affect their (already high) propensity to give help. Interestingly, we find that help increases across time at similar rates across treatments (likely because participants experience the performance benefit of receiving help similarly), which suggests that the effectiveness of goal setting and symbolic rewards is mainly due to an increase in the "baseline" help in the first round of the social media stage. This is also in line with results on the time trends for help presented in Section A.5 of the appendix, which generally show that participants that do (not) request or give help in one round continue to (not) request or give help in the following round.

2.4.3 Mechanisms of Helping Behavior

To explain the (non)effect of our treatments on outcomes, we look at the behavioral mechanisms that each treatment may be tapping into. Specifically, we look at the role of descriptive social norms in the [AI] treatment and reciprocity in the [FI] treatment. Since our badges treatment has a dramatically positive effect on helping behavior compared to both information treatments, we run additional treatments in Study 2 to elicit the behavioral mechanisms driving its effectiveness.

[1] In Section A.2 of the appendix, we also test for differences in canceling behavior for our additional treatments, where time to cancel was recorded, and find no significant difference across treatments.

Table 2.4: Helping Behavior and Rank

VARIABLES	w/ Individual Ability
AI Treatment	-0.406**
	(0.203)
FI Treatment	-0.0610
	(0.241)
Individual Ability	0.0390***
	(0.0109)
Previous Help	0.622***
	(0.0322)
Indicator Basic & Previous Rank is Low	0.223
	(0.224)
Indicator AI & Previous Rank is Low	0.871***
	(0.211)
Indicator FI & Previous Rank is Low	0.385*
	(0.231)
Constant	-0.0897
	(0.213)
Observations	816
Number of Participants	204

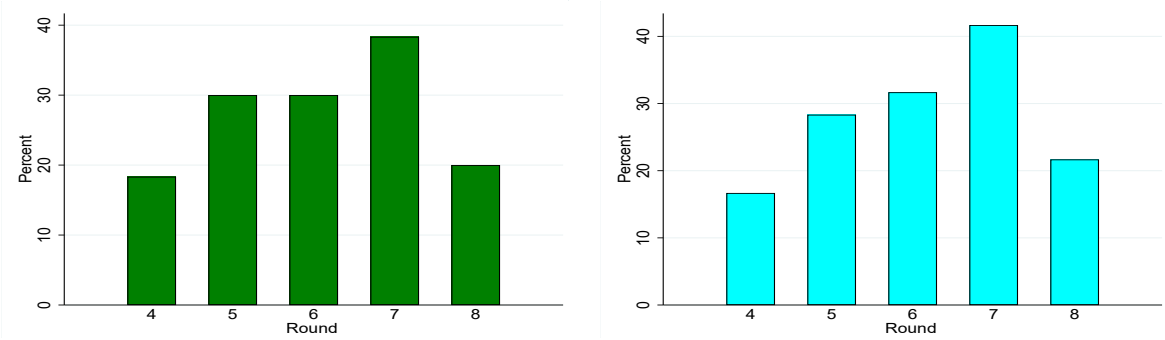
Robust standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

Notes: This is an OLS regression. The dependent variable is the amount of help given by an individual in a round. A helpfulness rank of 1, 2, or 3 is considered high rank and a rank of 4, 5, or 6 is considered low rank.

2.4.3.1 [AI] and Descriptive Social Helping Norms.

The [AI] treatment shows participants the total amount of help given and requested on the platform and their own helpfulness rank with respect to the group. This may encourage participants to change their helping behavior to conform to the group's descriptive helping norm. Table 2.4 presents the results from regressing the amount of help given by a participant in one round on treatment dummies and indicator variables for participants that had a low rank in the [AI], [FI], or [Ø] treatments. Unsurprisingly, we find no effect of previous rank on helping behavior in the [Ø] treatment, where no information on helping behavior is provided. However, our results show that, in the [AI] treatment, lower rank individuals help more and higher rank individuals help less. Thus, adding information on helpfulness

Figure 2.2: Percent of Participants Earning Each Badge in the [B] treatment.



(a) % Earning Early Help Badge

(b) % Earning Quantity of Help Badge

rank and aggregate helping does influence behavior in the [AI] treatment relative to the [Ø] treatment. However, it leads to an "averaging" of helping behavior within a group that results in the similar total quantity of help seen in our previous analysis. The lack of effect in the [FI] treatment suggests that participants may be more focused instead on individual helping information and reciprocity, which we discuss in the next section. We analyze helping behavior in the [AI] treatment further in Section A.6 of the appendix.

2.4.3.2 [FI] and Reciprocity.

The [FI] treatment builds on the [AI] treatment by adding information on each individual's helping information and the IDs of participants that answered a participant's help requests. This can help participants to target their limited amount of help, causing unhelpful individuals to wait longer or have their request remain unanswered. However, we find no systematic trends in either wait time or probability of receiving help by helpfulness rank (non-parametric test of trends, $p > 0.2$ for both). We also compare wait times across helpfulness ranks in the last 30 seconds of a round, where participants are pressed on time and may be more likely to target help, and also find no significant differences. In the appendix, we also find no evidence for participants targeting individuals helpful to them or targeting the most helpful individual over the least helpful individual in the group. We conclude that, in our experiment, participants in the [FI] treatment do not use the information available to them target their help towards helpful participants or participants helpful to them, contributing to the treatment's lack of effectiveness. We analyze helping behavior in the [FI] treatment further in Section A.7 of the appendix.

Table 2.5: Study 2 Behavioral Mechanism Treatment Comparison

Behavioral Mechanisms	$[\overline{\emptyset}]$	$[\overline{G3}]$	$[\overline{B3}]$	$[\overline{BE}]$	$[\overline{BE5}]$	$[\overline{BE35}]$
Goal Setting	✓	✓				
Symbolic Rewards		✓	✓			
Early Helpfulness Nudging			✓	✓		
Intermediate Goals and Feedback					✓	✓

Notes: The check marks show the relevant treatments used to disentangle each behavioral mechanism. G and B refer to goal-only and goal and badge treatments respectively, while E, 3, and 5 respectively refer to the presence of an "early help," "help 3 times," and "help 5 times" goal in a treatment. The $[\overline{\emptyset}]$ and $[\overline{BE5}]$ treatments are our original [Ø] basic platform treatment and [B] badges treatment repeated in our new online setting.

2.5 Study 2: Behavioral Mechanisms Driving the Badges Treatment

Since our badges treatment has a dramatically positive effect on helping behavior compared to both information treatments, we run additional treatments to elicit the behavioral mechanisms driving its effectiveness. Participants in the [B] treatment can complete an "early help" goal by helping once and requesting help once in the first 2 minutes of a round and a "quantity of help" goal by helping 5 times in a round. As seen in Figure 2.2, participants continue to earn both badges across the five rounds of the social media stage, suggesting that the badge rewards continue to be motivating to subjects even with repetition and a lack of monetary rewards. As discussed in our review of the literature, the effectiveness of our badges treatment could be due to a combination of goal setting, symbolic rewards, early helpfulness nudging, and/or intermediate goals and feedback. The design of our Study 2 treatments is intended to explore the effectiveness of each behavioral mechanism separately.

2.5.1 Experimental Design

Because of COVID-19 restrictions on in-person lab experiments, all treatments are done remotely. The treatments are run online using zTree Unleashed, with participants using a web browser. Participants are put in a Zoom meeting with other participants, asked to have their video camera on while doing the experiment, and are not allowed to talk or send messages to other participants. Similar to the physical setting, participants can see all other participants, but do not know which participants are in their group or what actions a specific

person is taking. Our methodology for running online experiments is similar to that used in Li et al. (2021). Our participants are given the same task and experimental setup as before. 414 undergraduate students participated in our new treatments. The interface of the online zTree experiment was smaller (due to it opening within a web browser tab) and its responsiveness was slightly slower than in the lab (possibly due to server performance, varying internet speeds, and/or increased latency compared to networked computers). Because the look and responsiveness of the platform are different with the online format, we repeat our basic platform and badges treatments online, referring to them as $[\overline{\emptyset}]$ and $[\overline{BE5}]$, and use them as reference in this section.

Table 2.5 provides a map of the experiments we conduct and their objectives. We first study the effect of goal setting alone on helping behavior by running a goal-only $[\overline{G3}]$ treatment that sets a "help 3 times" quantity of help goal that encourages participants to be helpful, but does not reward them with a badge. If goal setting contributes meaningfully to the effectiveness of the badges treatment, we should see a significantly higher quantity of helping behavior in the $[\overline{G3}]$ treatment compared to the $[\overline{\emptyset}]$ treatment. To examine the effect of symbolic rewards, we run a $[\overline{B3}]$ treatment that sets the same helping goal as the $[\overline{G3}]$ treatment, but rewards participants with a badge, possibly increasing their intrinsic motivation to pursue and complete the helping goal.

To separate the effect of goal setting and symbolic rewards from early helpfulness nudging, we run an "early help" only badges $[\overline{BE}]$ treatment and compare it to the "quantity of help" only $[\overline{B3}]$ treatment. The $[\overline{BE}]$ treatment rewards participants with a badge for helping once and requesting help once in the first two minutes of a round. While both treatments draw on goal setting and symbolic rewards to encourage participants to generally "be more helpful," the $[\overline{BE}]$ treatment also encourages participants to start using the platform earlier in the round. This gives participants more time to use the platform and can help establish helping norms earlier on. If early helpfulness nudging significantly contributes to the effectiveness of the badges treatment, we expect participants in the $[\overline{BE}]$ treatment to start requesting and giving help earlier in the round and to achieve higher quantities of help by the end of the round compared to the $[\overline{B3}]$ treatment.

Finally, since gamifiaction emphasizes the importance of having a structured sequence of goals, we study the effect of intermediate goals and feedback by running a $[\overline{BE35}]$ that adds an intermediate "help 3 times" quantity of help goal to the "help early" and "help 5 times" goals in the $[\overline{BE5}]$ treatment. Gamification principles suggest that this intermediate goal can give participants a chance to complete an easier quantity of help goal and receive positive feedback earlier, potentially increasing their intrinsic motivation to pursue the more

difficult "help 5 times" goal and improving their overall helping behavior. If intermediate goals and feedback significantly contributes to the badges treatment, we expect to see more helping behavior in the $[\overline{BE35}]$ compared to $[\overline{BE5}]$ treatment.

2.5.2 Participant Performance and Helping Behavior

Our results show that all our goal and badges treatments significantly improve performance over the $[\overline{\emptyset}]$ treatment. As in our previous section, we run a panel regression on performance in Section A.8 of the appendix. We find that, when accounting for differences in individual ability, all treatment coefficients are significant for all treatments. Furthermore, all treatment coefficients have reduced significance when we account for help requested in Column 3 or help received in Column 4, indicating that help requested/received is the driver of performance. We also find that helping behavior in all badges treatments is significantly higher than the $[\overline{\emptyset}]$ treatment (ranksum test, p>0.05 for all) and that helping behavior is marginally higher in the $[\overline{G3}]$ treatment compared to the $[\overline{\emptyset}]$ treatment (ranksum test, p=0.07). This indicates that, as in the previous section, improvements in performance can be largely attributed to increases in helping behavior and that introducing badges and helping goals significantly improves helping behavior.

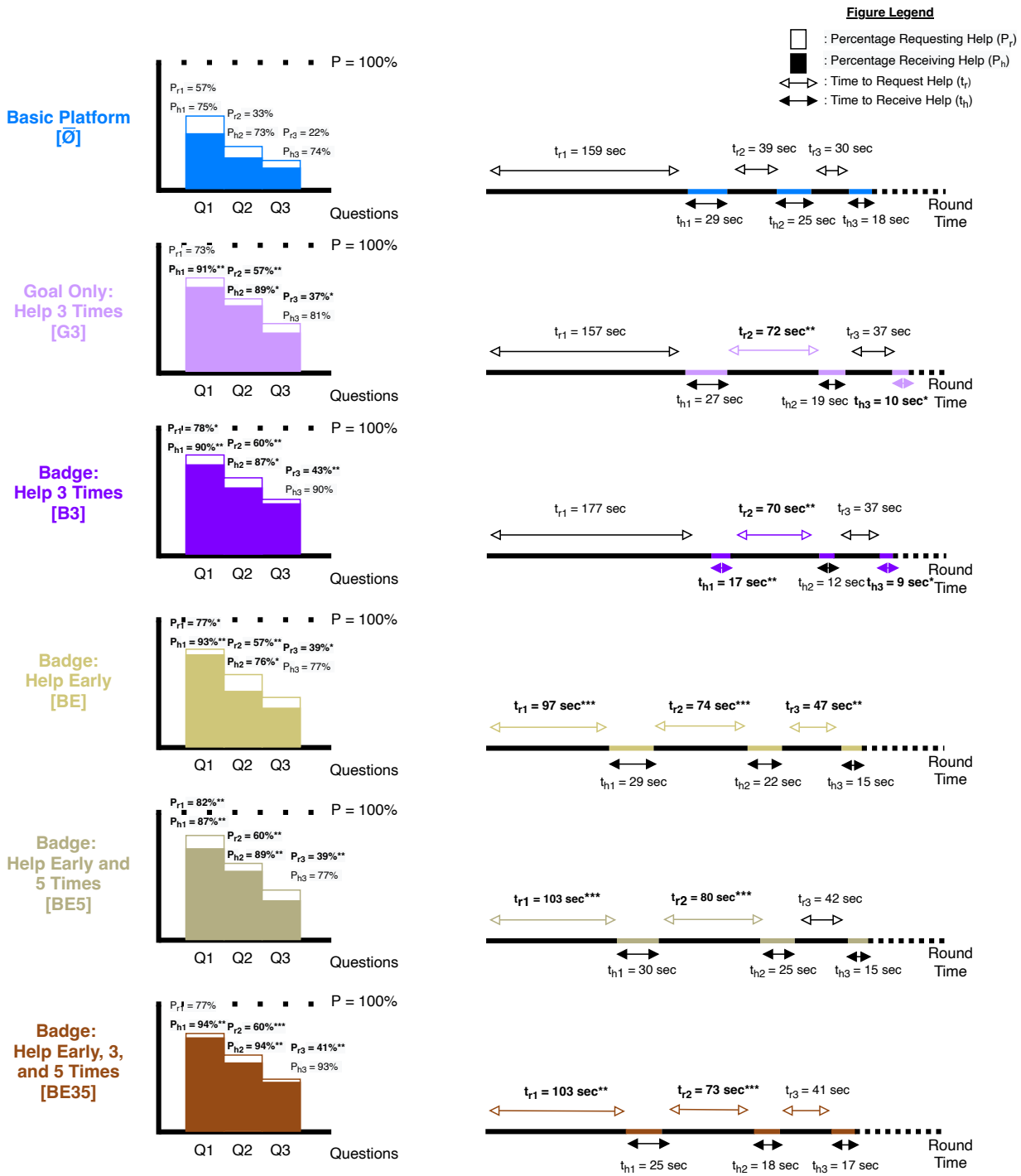
2.5.3 Mechanisms of Helping Behavior

To explore the effect of goal setting, symbolic rewards, early helpfulness nudging, and intermediate goals and feedback, we compare the quantity and dynamics of helping behavior across separate treatment pairings. As in the previous section, we calculate the percentage of participants requesting help, time to request, percentage receiving help, and time to receive help for the first three requests across treatments in Figure 2.3 and use them in our analysis of behavioral mechanisms.

2.5.3.1 Effect of Goal Setting.

As previously mentioned, $[\overline{G3}]$ treatment results in a marginal increase in helping behavior compared to the $[\overline{\emptyset}]$ treatment. The $[\overline{G3}]$ treatment also results in a significantly higher percentage of participants receiving help and a surprisingly higher percentage of participants requesting help, see Figure 2.3. Our results suggest that setting a helping goal that asks participants to give other participants help, even at the cost of leaving them with less time to work on their own tasks, is sufficient to encourage participants to both request and give more help on the platform, which leads to a significant improvement in their overall performance.

Figure 2.3: Helping Dynamics.



Notes: For each treatment, the percentage of participants that request help (P_r) and the percentage of participants that receive help (P_h) are shown on the left and the average time to request (t_r) and receive help (t_h) are shown on the right for the first three requests (denoted with subscripts 1 to 3) across treatments. Treatment comparisons are compared to the [$\emptyset$] treatment and are shown in bold text if significant: *** p<0.01, ** p<0.05, * p<0.1.

2.5.3.2 Effect of Symbolic Rewards.

Our results show that adding symbolic rewards in the $[\overline{B3}]$ treatment leads to a marginal increase in helping behavior over the $[\overline{G3}]$ treatment, with an average help given of 2.28 vs 1.77 (ranksum test, p=0.07). Furthermore, we find that the number of participants that would have completed the help 5 times goal had it been offered is marginally higher in the $[\overline{B3}]$ treatment than the $[\overline{G3}]$ treatment. We also find that the percentage of participants giving help and percentage of participants requesting help is greater in the $[\overline{B3}]$ treatment than the $[\overline{G3}]$ treatment, but the difference is not significant. Our results suggest that adding rewards, even ones that are both private and symbolic, can help improve helping behavior, resulting in added benefits to helping goals and in larger improvements in performance over the $[\overline{\emptyset}]$ treatment.

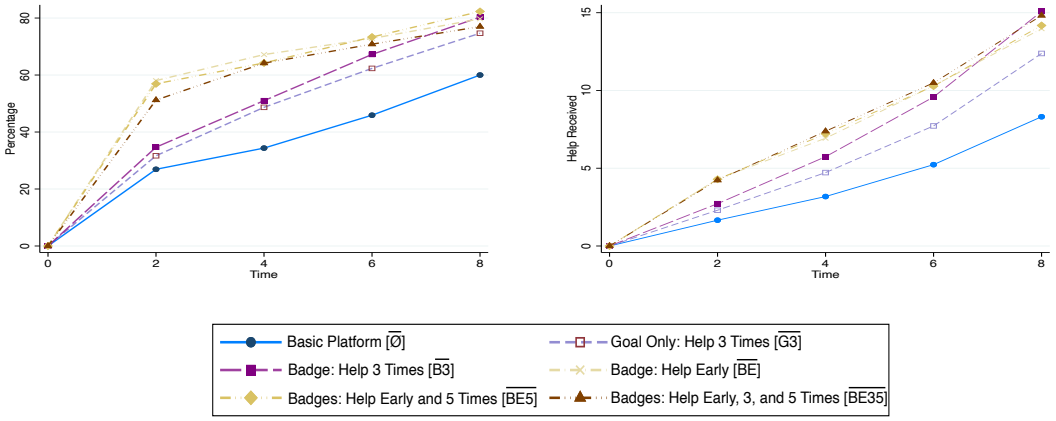
2.5.3.3 Effect of Early Helpfulness Nudging.

Figure 2.4(a) and Figure 2.4(b) respectively plot the average percentage of participants that request help at least once from the start of a round and the total amount of help received from the start of a round. Our results show that the "early help" $[\overline{BE}]$ treatment shifts help earlier in the round compared to the $[\overline{B3}]$ treatment, but that the two treatments result in the same quantity of help by the end of the round. Specifically, a significantly higher percentage of participants request help in the first two minutes of a round in the $[\overline{BE}]$ compared to the $[\overline{B3}]$ treatment (58.05% vs 34.67%, ranksum test, p=0.04). We also find significantly more participants earning the early badge in the $[\overline{BE}]$ treatment compared to those that would have earned it in the $[\overline{B3}]$ treatment had it been available (35.56 vs 12.33, ranksum test, p=0.00). Despite this large difference in helping behavior early on, the $[\overline{BE}]$ and $[\overline{B3}]$ treatments result in a similar quantity of helping behavior by the end of the round (2.07 vs 2.28, ranksum test, p>0.2). Surprisingly, our results suggest that early helpfulness nudging is not a major driver of helping behavior in our badges treatments, affecting only the timing but not the amount of help..

2.5.3.4 Effect of Intermediate Goals and Feedback.

Our results show no significant difference between the $[\overline{BE35}]$ and $[\overline{BE5}]$ treatments in the amount of help requested or percentage of participants requesting help and that both treatments result in almost overlapping graphs in Figures 2.4(a) and 2.4(b). This suggests that providing intermediate goals and feedback is also not a major driver of helping behavior in our badges treatments. Combined with our previous results on early helpfulness nudging, we conclude that gamification is not likely to be major factor behind the badges treatment's

Figure 2.4: Within Round Helping Pattern



(a) Percentage Active Participants on the Platform

(b) Amount of Help Requested

effectiveness.

To summarize, our Study 2 treatments help us elicit the behavioral mechanisms likely driving the effectiveness of our Study 1 badges [B] treatment. We find that the effectiveness of the badges treatment is primarily driven by goal setting and symbolic rewards that encourage participants to be helpful, rather than game-like mechanics in the form of early helpfulness nudging and intermediate goals and feedback.

2.6 Discussion

ESMPs can improve employee productivity by connecting them to coworkers capable of helping them. However, time-constrained employees may be concerned with the time it takes to use the platform, which may leave them with insufficient time to work on their own tasks. Our experiment focuses on such situations by constraining participants on time, incentivizing them based on their performance on a set of real effort tasks, and by giving them a platform that they spend their limited time on to request and give help to other participants. There are a number of ESMPs, with varying design features, that companies can choose from, but the existing literature provides little guidance on what design features are most effective. Our paper aims is to help companies identify specific ESMP design features that can help improve their employees' helping behavior and performance. In Study 1, we study the effectiveness of separate ESMP design features on helping behavior and performance and, in Study 2, we

identify the behavioral mechanisms driving the effectiveness of the most successful design feature from Study 1.

In Study 1, we run two information treatments that vary the level of information on platform helping behavior that is visible to a participant and find that neither significantly improves helping behavior or performance. We find that the [AI] treatment, which provides participants with information on total group helping behavior and their own helpfulness rank, results in unhelpful participants helping more but, interestingly, results in helpful participants helping less. While an averaging of contributions is often seen in public goods experiments where contributions are relatively costly and can drastically affect performance (see Keser and Van Winden, 2000), the result is surprising in our setting given the already small amount of help given (and time penalty incurred) by helpful participants.

Our [FI] treatment, which gives participants information on each group member's helping behavior, does not result in any targeting of helping behavior. While participants can observe the helpfulness of others (and target help accordingly), they do not know the performance of others. This may make them more reluctant to target help away from seemingly unhelpful participants, who may be struggling to answer their own questions and have little time to help others. Such "reluctance" to punish others is seen in public goods experiments where contributions are public, punishment is allowed, but participants' initial endowment is kept hidden (Bornstein and Weisel, 2010).

Our [B] treatment, surprisingly, results in a significant increase in helping behavior and performance. The [B] treatment sets one "early help" goal and one "quantity of help" goal and rewards participants with a private and symbolic badge for completing each goal. Participants that complete the "early help" goal also make progress towards completing the "quantity of help" goal. As such, the effectiveness of the [B] treatment could be due to goal-setting motivation, symbolic reward attainment, early helpfulness nudging, and/or intermediate feedback. In Study 2, we vary the design features of our [B] treatment, and show that the treatment's effectiveness is due to both goal-setting motivation and symbolic reward attainment, which encourage participants "to be more helpful," rather than game-like mechanics that seek to alter the way they use the platform and/or that provide them with intermediate feedback and encouragement.

Our results speak to the effectiveness of platform design features that leverage a participant's intrinsic motivation to help. The helping goals we set and the private symbolic rewards we award in the [B] treatment ask participants to exert costly effort without directly increasing their monetary payoff and without affecting their public image or reputation. Specifically, participants in our experiment are constrained on time and are paid solely on

the number of word puzzles they answer correctly, yet our helping goals ask them to spend their limited time helping others. Despite this "apparent" misalignment between incentives and goal requirements, participants accept and pursue our helping goals and the result is a higher percentage of participants giving and requesting help and in participants requesting and giving help more frequently. Furthermore, we find that private symbolic rewards, in the form of private badges, are effective at improving helping behavior even though their primary benefit is only towards a participant's own self-image. Our results can help explain why popular platforms such as Facebook Workplace and Jive choose to allow employees to earn badges for completing certain goals on the platform and why other platforms such as Microsoft Yammer should consider adding badges to their platform design.

2.7 Conclusion

Based on the results of our paper, companies interested in implementing ESMPs might consider leveraging their employees' intrinsic motivation to help by setting helping goals on the platform that employees can complete by giving help and requesting help on the platform and by rewarding employees that complete these goals with non-monetary rewards, that could possibly be private in nature, as recognition of their effort and achievements. While we find no evidence of reciprocal behavior in our experimental setting, this behavior may differ in real-world settings where there are likely many more factors affecting helping decisions. Our experiment also finds that providing helping information that possibly leverages descriptive social norms may harm helping behavior by encouraging participants that help more than average to help less.

While we focus on employees constrained on time who worry about the time burden of using ESMPs to request and give help, employees can have other reservations, such as the fear of losing their value once their knowledge is made public on the platform. Future research could study how incentive mechanisms can be designed to reassure employees about sharing their knowledge on the platform. Participants spend a limited amount of time participating in our experiment and are anonymously paired with other participants. Future research should study the effect of social norms, reciprocity, and goal-setting and symbolic rewards in real-world platforms where employees are likely to know one another and interact over a longer period of time. Our paper also assumes that employees can easily ask questions and receive correct answers. Future research could study helping behavior in situations where problems are sophisticated, which increases the difficulty of asking and answering them correctly and makes receiving an incorrect answer more likely.

CHAPTER 3

Creative Task Constraints and Knowledge Worker Productivity

3.1 Introduction

Knowledge workers often work on creative tasks that involve elements of both originality and usefulness (Hopp et al., 2009; Amabile et al., 2018). A creative output's originality is broadly measured based on how different it is from what is normally seen for that task and its usefulness is measured based on its ability to "create value." In many work settings, the usefulness of a knowledge worker's idea can depend on with how understandable it is to a separate set of employees tasked with implementing it (Ostermaier and Uhl, 2020). In such settings, a knowledge worker can be asked to create output that is original and that meets a certain level of usefulness for the company. For example, an architect at a small architectural firm can be tasked with creating novel designs that can win over the client (originality goal) and that can, with varying degrees of effort, be understood by the firm's construction team tasked with implementing the idea (usefulness constraint).[1] A trade-off could then exist between submitting an original idea that the client and, likely, the construction team have never seen before and a useful idea that the construction team can understand and that they and, likely, the client are familiar with.[2] Our paper focuses on such knowledge worker settings where a trade-off between an idea's originality and its usefulness. In such cases, varying usefulness constraints can cause an employee to under-emphasize or over-emphasize usefulness, resulting in output that might be original but not sufficiently useful or that is exceedingly useful but not very original. We run a lab experiment to answer the following

[1]For example, a unique building design that the construction team cannot understand is likely to not be useful for the firm, and, as such, would not be considered creative.

[2]This trade-off between originality and usefulness can exist in other knowledge worker settings as well. For example, a user experience engineer may want to create an exciting design that users have never seen before (originality) and one that they can understand and, thus, use (usefulness).

questions: How do usefulness constraints affect employee performance in an originality-focused creative task? Can managers benefit from artificially tightening or relaxing the usefulness constraints they set for employees?

Our paper is the first to study usefulness constraints, a form of output constraints, in a lab setting. Much of the constraints literature has focused on constraints on the input or process of a creative task (Kagan et al., 2018; Moreau and Dahl, 2005). With input and process constraints, participants are always forced to factor in the constraint in their creative output. For example, a participant may have fewer building materials to work with or may have less time to work on a creative task. With usefulness constraints, it is up to the participant to decide if and to what extent they want to factor in the constraint (they can also choose to ignore the constraint completely). Similarly, some output constraints can be in the form of "physical" requirements, such as asking a participant to include certain combinations of building materials in their output. In such cases, a participant can immediately verify if their creative output meets the requirements. This is not the case with usefulness constraints, where a participant is unlikely to know the exact usefulness of their creative output.

Our experiment asks participants to work on a creative task where they are incentivized based on the originality of their creative output on condition that it meets a usefulness constraint. Participants are tasked with creating eight original images using a set of drawing materials to create an object and a set of emoji to represent an action on the object. We associate an image's originality with how different it is compared to other images in the experiment. Originality is measured by a set of research assistants that have been trained on a relatively large number of practice images from the experiment. As with our architect example, a creative idea's usefulness or implementability can depend on how easily it can be understood. We measure an image's usefulness in our setting by capturing how "recognizable" to a set of raters. Specifically, an image's recognizability is calculated as the number of raters able to guess its exact noun by only seeing its emoji, object, and verb. We then vary the usefulness constraint across four treatments to study the effect that has on the originality and usefulness of a knowledge worker's creative output and on their overall performance. Our experiment consists of four treatments. Our T0 treatment sets a 0% constraint, effectively paying participants based on the originality of their creative output regardless of its usefulness. We then set "low," "moderate," and "high" usefulness constraints in the T10, T40, and T80 treatments by setting the usefulness constraint of 10%, 40%, and 80% respectively. Our treatments are designed to encourage participants to focus on originality in the T0 treatment, to focus on originality while accounting for a minimal level of usefulness in the T10 treatment, to focus on both originality and usefulness in the T40 treatment, and to focus primarily on usefulness in the T80 treatment.

We find that, as expected in our setting, a trade-off exists between an image's originality and its usefulness. Our results show that adding a low usefulness constraint in the T10 treatment results in images that are significantly more useful and significantly less original than the T0 treatment. This results in participant performance that is higher when accounting for the 10% usefulness requirement. Surprisingly, adding a moderate usefulness constraint in the T40 treatment generates images with the same usefulness as the T10 treatment, but with significantly worse originality. This results in participant performance that is significantly worse in the T40 treatment compared to the T10 treatment when accounting for the 40% usefulness requirement. Finally, adding the high usefulness constraint in the T80 treatment generates images with a similar originality and usefulness as the T10 treatment and that, importantly, mostly fail to meet the 80% usefulness constraint. This results in participant performance that is exceptionally poor and that can be improved by even a rudimentary strategy of submitting eight basic recognizable images.

Our findings suggest that low usefulness constraints can improve employee performance by an effectively "nudging" them to factor in usefulness as they pursue an originality-focused creative task. Interestingly, we find that moderate and high usefulness constraints can be ineffective or even detrimental to employee performance, causing employees to either generate output that is creatively poor or to fail to properly account for the usefulness constraint. Our paper suggests that managers, in such cases, can improve employee performance by "artificially" lowering the constraint they set for their employees or by possibly changing the goal of the creative task towards usefulness rather than originality.

3.2 Literature Review

3.2.1 Knowledge Worker Productivity in Operations Management

Performance and productivity have been long-standing topics of interest in operations management (Smith and Robey, 1973; Ebert, 1976; Fujimoto and Clark, 1991; Herroelen and Leus, 2005; Schmenner, 2015). Much of the productivity research in operations has focused on improving productivity in settings traditionally associated with standardized work, where tasks are often physical and repetitive in nature, sometimes referred to as blue-collar work. This research considers topics such as the effects of work sharing, individual and group incentives, task switching, inventory policies, and queue structure on productivity (Schultz et al., 1999; Shunko et al., 2018; Stratman et al., 2004; Bendoly et al., 2014; de Vries et al., 2016). As opposed to production, service and professional jobs are typically more creative and knowledge intensive, are inherently less certain than physical tasks, and give employees

greater discretion in carrying out tasks based on their judgment and learning capabilities (Spohrer and Maglio, 2008). In our paper, we focus on settings where a knowledge worker is pursuing an originality goal subject to a usefulness constraint. In the knowledge worker settings we consider, the usefulness of a knowledge worker's idea is associated with how understandable it is to a separate set of employees likely tasked with implementing it. For example, a knowledge worker that submits a unique design that other employees find difficult to understand and then implement could be considered less useful and, as such, less creative than a unique design idea that employees can easily understand. While the knowledge worker knows the specific description of her assigned task, she does not know beforehand if and to what extent her creative output is original and useful and if it meets the usefulness constraint. Our goal is to study how a manager can use and/or adjust usefulness constraints to productively influence the originality and usefulness of a knowledge worker's output and her overall performance.

3.2.2 Creativity Literature

Creative ideas are often defined as those that are both original and useful (Amabile et al., 2018). The originality of an idea is often judged based on its uncommonness or uniqueness and its usefulness is often measured based on its ability to generate utility or "create value" in the domain or context in which it exists (Runco and Jaeger, 2012). While the measure of originality is relatively similar across domains, the specific definition and measure of usefulness is often context dependent. For example, Berg (2014) asked college students to create original and useful product ideas for the university's bookstore. The originality and usefulness of ideas were judged by bookstore managers and customers based respectively on how different a product is from what exists at the bookstore and in general and on its propensity to create value. In a separate setting, Ostermaier and Uhl (2020) asked participants to brainstorm words that other participants can then use to write a piece of text. A word's originality was measured based on how uncommon it is compared to other submitted words and its usefulness was measured based on how many participants then selected it to use in their writing. In our experimental setting, we associate an originality of a participant's submitted image with its uncommonness compared to other submitted images in the experiment (as determined by our trained research assistants) and its usefulness with how understandable, i.e. recognizable, it is to a separate set of participants (raters on Prolific).

The creativity literature suggests that it is relatively difficult to generate ideas that are both original and useful and that, as a result, a trade-off often exists between the two (Oster-

maier and Uhl, 2020; Berg, 2014; Miron-Spektor and Beenen, 2015). In our experiment, we ask participants to create original ideas that meet a recognizability constraint and vary the recognizability constraint across treatments (from 0% in T0 to 10% in T10, 40% in T40, and 80% in T80). Participants in our experiment are paid based on how different their *image* is compared to other images in the experiment, which depends on the image's idea and its implementation. This means that, even in the T0 treatment, we expect a participant pursuing an original idea to likely leverage details unique to her image's idea in its implementation to make it "stand out" from the images submitted by other participants.[3] As such, we expect that participants, even in T0 treatment, will factor in some form of recognizability in the implementation of their images.

We expect that attempting to factor in a higher recognizability constraint will generally increase the emphasis participants place on recognizability, increasing the recognizability of their images and decreasing their originality. Our goal is to see if and how participants factor in the recognizability constraints we set and to study the magnitude of change that has on the originality and recognizability of their creative output. If a constraint causes participants to under-invest or over-invest in recognizability or to generally have a worse originality-recognizability trade-off, then participants might be better off pursuing variations of the constraint that either encourage them to place more accurate emphasis on recognizability or to have a better originality-recognizability trade-off.

3.2.3 Experimental Literature on Constraints

The literature examining the effect of constraints on creativity can generally be divided into input, process, and output constraints (Acar et al., 2019). Input constraints often limit the resources, such as materials and time, that participants have access to when working on a creative task and process constraints limit the autonomy or how much freedom participants have when approaching a creative task. The literature finds that input and process constraints, if implemented correctly, can improve the creativity of a participant's creative output. For example, Scopelliti et al. (2014) asked participants to create a toy design from a set of building materials and showed that setting input constraints by limiting the number of materials participants can use can improve the creativity of the designs they generate. In a separate setting, Kagan et al. (2018) show that participants that work on a creative task consisting of an ideation phase and an execution phase perform better when the switching

[3]For example, if a participant uses one empty circle to represent "Mars", the idea of creating the planet Mars would be unique in our setting, but its implementation would be similar to the many other images containing one circle (such as "plate") submitted by other participants. The judges rating the image would likely give such an image lower points on overall originality than an image using a red circle to represent Mars and circles to represent craters for example.

point from ideation to execution is imposed on the participant (a form of process constraint) rather than when it is left up to the participant to decide.

Our paper focuses instead of output constraints, in the form of usefulness constraints. While input and process directly change how a participant can approach the task (for example, a participant starts the task with less building materials), output constraints rely on a participant's decision to incorporate them. To our knowledge, one other paper has studied output constraints in an experimental setting. Moreau and Dahl (2005) place output requirements by asking participants to submit toy designs that use all five building materials given to them (rather than designs that use any number of the building materials). They show that adding such an output constraint can improve the creativity of a participant's creative output. In our paper, we focus specifically on usefulness output constraints. Importantly, participants in our setting do not know if their creative output meets the usefulness output constraint we set when they submit their creative output. As such, "artificially" relaxing or tightening the usefulness constraint, which we examine in our paper, can improve participant performance by helping them either be more original and/or helping them submit more useful output that helps them meet their assigned usefulness constraint.

3.3 Experimental Design

Our experiment consists of two parts. In Part 1, participants work on a creative task that incentivizes them based on the originality of their creative output on condition that it passes a usefulness constraint. In Part 2, we use a survey to measure a participant's risk aversion and to record how they approached the experiment and their beliefs on the originality and usefulness of their creative output. We run four treatments that steadily increase the level of usefulness required to meet the constraint. This allows us to measure the effect that increasing usefulness constraints have on the originality and usefulness of participant's creative output and on their overall payoff.

Our creative task asks participants to create images using a set of drawing materials and emoji. For each image, a participant is asked to use the building materials to create an object, which they place on the right of the image, and one emoji to depict an action on that object, which they place on the left of the image. For example, a participant can use the drawing materials to create a car as their object and the "bag of money" emoji to create a "buy car" image (see Section B.2 of our appendix for examples). Participants are given thirty minutes to create up to eight images. We set these parameters so that participants across all treatments generally have a plentiful amount of time to submit all eight images. This allows us to focus on how the originality and usefulness of a participant's creative output varies

across treatments without worrying about changes in the quantity of images submitted.

Participants are given the same building materials to use for each image. The building materials are twelve colored pebbles (three red, three blue, three orange, and three green), four rectangular sticks, and two circles. This is inspired by the experimental task used in Laske and Schroeder (2017). Participants are allowed to change the dimensions and rotation of the building materials, but they cannot add any building materials or change any of their other attributes such as their color or transparency. The building materials are chosen to give participants a good amount of flexibility in the objects that they can create, while also reflecting the constraints on materials and budget that knowledge workers, such as architects, have to account for when they are generating designs. Participants are also given a set a eight emoji and they can only use each emoji once. The emoji are chosen to be relatively different from each other, while also being relatively flexible so that they can be used to represent a number of different actions (see Section B.1 of our appendix for the exact building materials and emoji). This is to reflect a situation where a knowledge worker, such as an architect, might have to create different categories of designs. Finally, after creating an image, a participant must write down a one word noun denoting the object they created and a one word verb denoting an action of the emoji on the object. Restricting the verb and noun to one word each makes it easier to measure originality and usefulness, which we discuss below.

We incentivize participants based on the originality of their submitted image on condition that it passes a usefulness constraint. To measure originality, we use the Consensual Assessment Technique (CAT) proposed by Amabile and Pratt (2016). Participants are informed that the originality of an image is measured by our judges based on how "different" it is compared to other images submitted in the experiment. The judges we use in our setting are trained research assistants (RAs). RAs use a rubric to rate each image across nine different criteria inspired by the CAT and to give each image an overall originality rating. Each image is judged by two RAs and the average of the two ratings is used for the final image rating for each criteria. Importantly, RAs are given a detailed description of the experimental task that participants work on, but do not know the exact purpose of our study, how our treatments vary, or which treatment each image belongs to. Our inter-rater reliability is fairly high for all dimensions (Krippendorff's $\alpha > 0.77$ for all dimensions). We refer the reader to Section B.3.1 of the appendix for more details about how we use the CAT in our setting.

We focus on settings where an idea's usefulness depends on its implementability (Ostermaier and Uhl, 2020). For example, an architect's design may be deemed less useful for the firm if the construction team finds the design difficult to implement. In our setting, we assume that an image that is difficult to understand is likely to not be useful or implementable.

As such, we measure an image's usefulness or implementability with how easily it can be understood by a set of inexperienced raters.[4] Participants, in our setting, are then paid based on the originality of their submitted image on condition that it passes a recognizability constraint. Participants are informed that their submitted images will be presented without their noun to a random set of raters on Prolific, an online crowd-sourcing platform, who are asked to guess the noun (based on the object, emoji, and verb). The percent recognizability of an image is the percentage of the raters that can correctly guess its exact noun. Each image is presented to exactly ten raters and Prolific raters are incentivized based on the number of nouns they guess correctly. We refer the reader to Section B.3.2 of the appendix for more details.

3.3.1 Participants

Two hundred and forty participants are recruited at the University of Michigan. The experiment was conducted online using Zoom and was implemented using Google Slides for the creativity task and Qualtrics to measure risk aversion and for the exit survey. Participants are paid a $5 show up fee and a $2 for completing the experiment in addition to the money they make from the creativity task and the risk aversion survey. Payments ranged from $9 to $24.

3.3.2 Procedure

Upon joining the Zoom room, each participant is privately given access to a separate Google Slides deck containing the instructions for the creativity task. Participants are asked to have their video camera on for the duration of the experiment and are not allowed to talk to one another. At the start of the experiment, the instructions in the slide deck are read out loud. Participants are then given a set practice questions to make sure they understand how their payoff is calculated, how the originality and recognizability of their images is measured, and the image criteria they need to follow (to avoid having their image disqualified). They are also given a practice slide with a ninth (mouth) emoji and asked to create a practice image on that slide. We then go through the practice questions together and check each individual's practice image to make sure it meets all the criteria and give them feedback to clarify any misconceptions. We do not give participants any practice images to avoid having them fixate

[4]We use inexperienced raters to capture an image's recognizability more accurately. While trained raters can guess that a design is a car if it is properly or poorly implemented, a set of inexperienced raters may have a harder time correctly guessing the poorly implemented car, allowing us to differentiate between the two implementations.

on any particular idea Berg (2014). We also choose the mouth emoji to be relatively flexible, but distinct from the other emoji they have in their set.

After answering any remaining questions, we paste the eight submission slides directly on each individual's slide deck and start the 30 minute timer for the creativity task. At the end of Part 1, we change the permissions on the slides deck to allow participants to view, but no longer edit, the slides so that they can refer to them when they answer our exit survey questions. We then send them a Qualtrics survey link that contains the risk aversion survey in Part 2 and the exit survey. After the experiment, participant images are collected and arranged into Qualtrics surveys that are sent to Prolific raters to measure recognizability and to separate Qualtrics surveys sent to RAs to measure originality. Each survey contains images from all four treatments placed in random order.

3.3.3 Treatments

Our experiment consists of four treatments that differ in the recognizability constraint they set. The T0 treatment sets a 0% recognizability constraint. This effectively incentivizes participants based only on the originality of the images they submit (regardless of their recognizability) and allow us to see how "original" original images can be and the baseline recognizability of these images in our experiment. We set the constraint at 0% rather than just removing the constraint so that participants in the T0 treatment are still aware of image recognizability and how it is measured (even if it does not factor into their final payoff), allowing us to directly compare the T0 treatment to our other treatments with higher recognizability constraints.

Our T10 treatment introduces a "low" recognizability constraint by increasing the recognizability requirement to 10%. This means that participants must submit an image that can be understood by at least one of the raters to receive payment for that image. We set this constraint to encourage participants to account for a minimal level of recognizability as they pursue an originality goal for their submitted images. Our T40 treatment increases the recognizability requirement to a "moderate" 40%, requiring that the image be understood by around half of the raters. This is designed to encourage participants to put emphasis on both originality and recognizability as they pursue an originality goal for their images. Finally, our T80 treatment sets the recognizability to a "high" 80%. This means that a participant must submit an image that is understood by almost all raters to receive payment for that image. We set this constraint to encourage participants to primarily emphasize recognizability despite the originality goal set for their images.

3.4 Results

The goal of our experiment is to study how a participant's performance on an originality-focused creative task varies across increasing recognizability constraints and if we can improve participant performance by "artificially" varying the recognizability constraint we set. As a reminder, participants in each treatment are paid based on the originality of each image they submit on condition that it meets the recognizability constraint of that treatment. We start by simply examining participant payoffs in each treatment (each under their own recognizability constraint). Then, for each recognizability constraint, we run a counterfactual analysis to see how participants in the other treatments would perform subject to that constraint. We find interesting and surprising effects on participant performance and show that participants can, in certain situations, benefit from pursuing artificially different task constraints. To explain our treatment results, we study the effect of different recognizability constraints on the recognizability and originality of participant's submitted images and, possibly, on how they approach our creative task.

3.4.1 Treatment Performance

In Table 3.1, we display, for each treatment, the average number of images passing the recognizability constraint, the average originality of images passing the constraint, and a participant's subsequent payoff on average.[5] Our results show that increasing the recognizability constraint expectedly leads to a significant decrease in a participant's payoffs (rank sum test, p=0.00 for all treatment comparisons) because it generally decreases both the number of images passing the recognizability constraint and the originality of images passing that constraint.

Our goal now is to see if a participant working under one recognizability constraint would benefit from pursuing an artificially different constraint. Figure 3.1 shows the performance of our four treatments across the four constraints. Our T0 treatment effectively encourages participants to only focus on the originality of the images they submit. As expected, we find that the T0 treatment results in the highest participant payoff at the 0% constraint, i.e. when we compare the originality of all submitted images (p=0.02, p=0.00, and p=0.01 compared to T10, T40, and T80). Our T10 treatment is designed to encourage participants to emphasize a minimal level of recognizability as they pursue an originality goal. Similarly, the T10 treatment results in the highest participant payoff at the 10% constraint, i.e. when

[5]We note that, as intended by our design, participants across the four treatments almost always submit all eight images and there are not significant differences across treatments (ranksum test, p>0.2 across all treatments).

Table 3.1: Summary Statistics

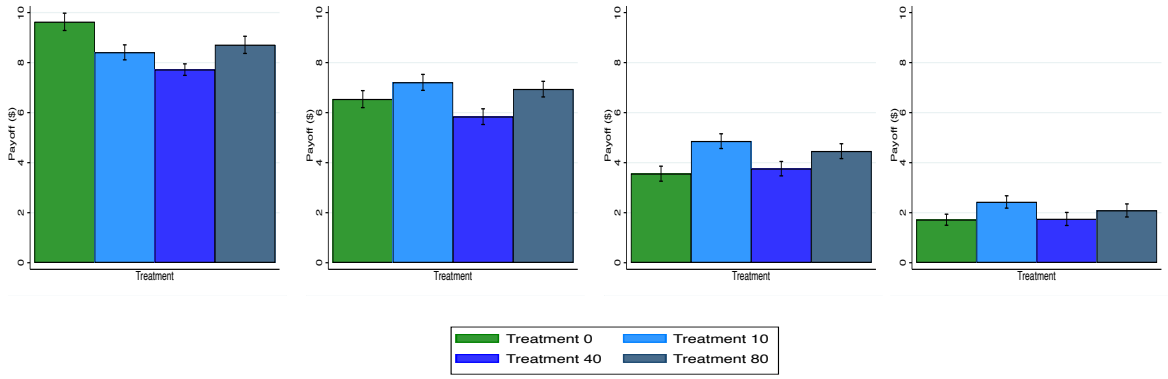
Treatment	Number of Images Passing Constraint	Originality of Images Passing Constraint	Payoff
T0	7.95	3.08	$9.63
T10	6.7	2.69	$7.22
T40	4.32	2.19	$3.78
T80	2.05	2.54	$2.09

we compare the originality of images that pass a 10% recognizability constraint (p=0.13 compared to T0, p=0.00 compared to T40, p>0.2 compared to T80).

We find surprising results when we look at participant payoffs at the 40% and 80% constraints. Our T40 treatment is designed to encourage participants to emphasize both originality and recognizability. Interestingly, the T40 results in an average participant payoff that is significantly worse than the T10 treatment at the 40% constraint (p=0.01 compared to T10, p=0.08 compared to T80, and p>0.2 compared to T0). This means that participants pursuing an originality goal with an "artificially" lower 10% constraint would receive a higher payoff than participants that pursue the "true" 40% recognizability constraint. Finally, our T80 treatment is designed to encourage participants to emphasize recognizability despite the originality goal for their creative task. We find that participants in the T80 treatment perform exceptionally poorly, receiving an average payment of $2.09 (performance is similarly poor across treatments, minimum p=0.19). Participants can substantially improve their payoff by submitting a relatively simple everyday object (such as a car) for all eight images. Doing so would result in the participant having eight images that pass the 80% constraint, which with a minimum originality rating of 1 out of 5 for each, would result in an average payoff of $3.2 at the 80% constraint.

Our goal now is to see why setting a "low" 10% constraint increases participant payoffs for that constraint, while setting a "moderate" 40% or a "high" 80% results in relatively poor participant payoffs for those constraints. We start by examining the effect of increasing the recognizability constraint across treatments on the recognizability and originality of a participant's submitted images. We plot average image originality and recognizability across treatments in Figure 3.2. Our results show that increasing the recognizability constraint, surprisingly, leads to both non-linear and non-monotonic changes in average image originality and recognizability. We now explore this further.

Figure 3.1: Performance Across Treatments



(a) 0% Constraint (b) 10% Constraint (c) 40% Constraint (d) 80% Constraint

Note: The above graphs show how much a participant in each of the T0, T10, T40, and T80 treatments would earn subject to each of the 0%, 10%, 40%, and 80% constraints.

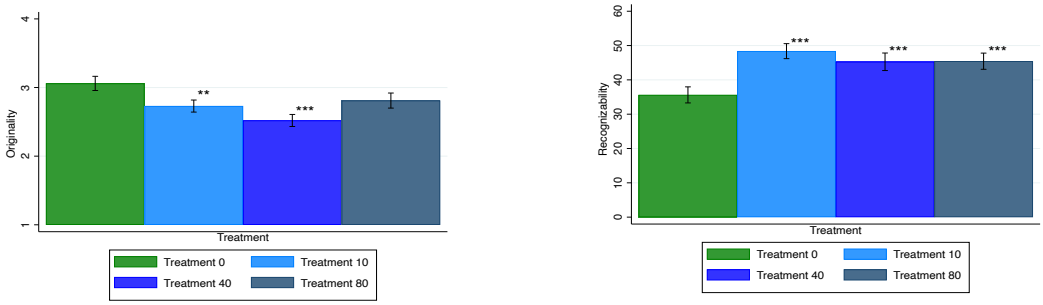
3.4.2 Treatment Performance

Participants in our experiment are paid based on the originality of each image they submit on condition that it meets a certain recognizability constraint. For example, a participant in the T10 treatment is paid for each image they submit that passes a 10% recognizability constraint. For the purpose of our analysis, we will study how participants in the T0, T10, T40, and T80 treatments perform if their submitted images have to meet each of a 0%, 10%, 40%, and 80% constraint. This allows us to see if participants that pursue one constraint can have perform better if they potentially pursued another higher or lower constraint. We note that, as intended by our design, participants across the four treatments almost always submit all eight images and there are not significant differences across treatments (ranksum test, $p>0.2$ across all treatments).

3.4.3 Image Recognizability Across Treatments

Recognizability, in our setting, is measured based on how many raters are able to guess the noun of an image the participant wrote down by only seeing the image's emoji, object, and verb. Our "baseline" T0 treatment incentivizes participants based on the originality of their images, regardless of their recognizability. This gives us a benchmark of an image's baseline recognizability if participants are pursuing an originality goal. We find that images in the T0

Figure 3.2: Average Image Originality and Recognizability



(a) Average Originality

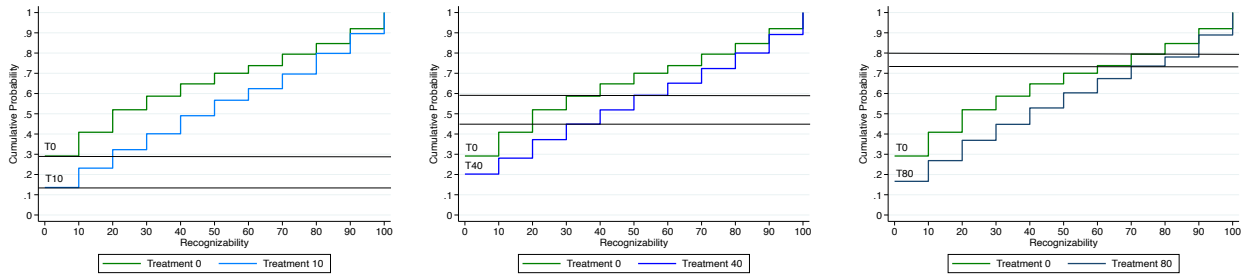
(b) Average Recognizability

treatment have an average image recognizability of 35.62%.[6] o see how many of the images in the T0 treatment are to likely to "naturally" meet the 10%, 40%, and 80% constraints, we plot a histogram of image recognizability in Figure 3.3. Our results show that, of the images submitted in T0, 71% pass a 10% constraint, 41% pass a 40% constraint, and only 20% pass an 80% constraint. This means that, as designed, our 10%, 40%, and 80% constraints correspond to "low," "medium," and "high" recognizability constraints. Specifically, despite the originality goal, participants in the T10 treatment likely need to place more emphasis on recognizability to avoid having a third of their images disqualified by failing to meet a 10% recognizability constraint. Similarly, participants in the T40 and T80 treatments would need to place increasingly even emphasis on recognizability to, respectively, avoid having more than half and almost all of their images disqualified. We now look at image recognizability in each treatment.

Our results show that the T10, T40, and T80 treatments result in significant, but similar, increases in recognizability over the T0 treatment (48.37%, 45.14%, and 45.55% vs 35.62%, ranksum test all p=0.00). As desired, the T10 treatment results in significantly more images passing the 10% constraint compared to the T0 treatment (87% vs 71%, p=0.00) and the T40 results in significantly more images passing the 40% constraint compared to the T0 treatment (55% vs 41%, p=0.00). Interestingly, the T80 does not result in significantly more images passing the 80% constraint (27% vs 20%, p=0.14). We note that, while only a small portion of images in the T80 treatment pass the 80% constraint, it is possible for participants to submit more recognizable images and ones that can easily pass the constraint.

[6]In Section B.4 of the appendix, we discuss why baseline recognizability might be relatively high in our setting.

Figure 3.3: Distribution of of Image Recognizability Across Treatments



(a) Treatment 10 (b) Treatment 40 (c) Treatment 80

Notes: We plot the cumulative density functions of image recognizability in the T10, T40, and T80 treatments with the T0 treatment as reference. The black horizontal lines in each figure show the percentage of images in each treatment that do not pass its respective constraint. The difference between the two black lines in each figure shows the increase in the percentage of images that pass the constraint over the T0 treatment (where the constraint was not present).

In a pilot of a recognizability-only (TR) treatment, we find that, compared to T80, images have a significantly higher average recognizability (60.3% vs 45.55%, p=0.01) and that significantly more images pass the 80% constraint (27% vs 42.5%, p=0.00). More generally, participants in any treatment of our experiment can repeatedly draw everyday objects such as "house," "car," or "flower" that almost always exceed the 80% constraint. Our results suggest setting low or moderate recognizability constraints is an effective way to encourage participants to submit images that pass those constraints more often. We find that setting a high constraint is a surprisingly ineffective way to encourage participants to submit images that pass that constraint.

We know that recognizability is the same across our three treatment, so our goal now is to see if originality is also the same across the three treatments. Specifically, the T10 treatment has the same image recognizability as the T40 treatment, but results in better participant payoffs subject to the 40% constraint. We now check to see if this is because the T10 treatment results in better image originality. Participants in the T80 treatment should be primarily emphasizing image recognizability, likely to the detriment of image originality. To see if that is the case, we also compare the originality of images in the T80 treatment to the other treatments.

3.4.4 Image Originality Across Treatments

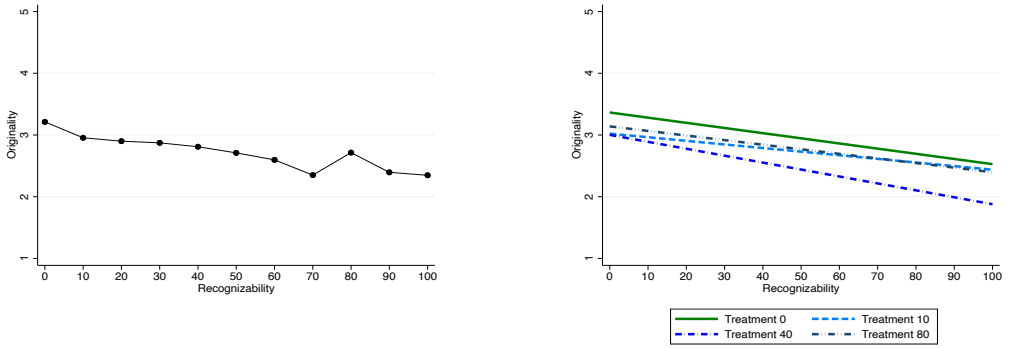
Image originality is measured by our judges based on how "different" an image is compared to other images submitted in the experiment. As mentioned previously, there is likely a trade-off between image originality and recognizability in our setting. As expected, the T10 treatment result in a significant decrease in average originality compared to the T0 treatment (3.06 vs 2.73, p=0.02). We get a similar results when we regress image originality across treatments, clustering for each participant's images in Table 3.2 Column 1. Taken together with our previous results on image recognizability, we find that introducing a low 10% constraint results in participants submitting images with higher recognizability and lower originality compared to the T0 treatment. This results in participant payoffs in the T10 treatment that are lower than the T0 treatment for the 0% and that are higher for the 10% constraint. Our results suggest that low recognizability constraints are an effective way to improve participant performance by successfully encouraging them to factor in more recognizability in their creative output.

We now compare the T10 and the T40 treatments and find that the T40 treatment surprisingly results in a significantly lower average image originality (2.73 vs 2.5, p=0.04). Taken together with our recognizability results, we find that the T10 treatment is able to generate output that is as recognizable as the T40 treatment, but that is significantly more original. This results in participant payoffs in the T10 treatment that are significantly higher than the T40 treatment for the 40% constraint. Our results suggest that moderate recognizability constraints can be detrimental to participant performance by causing participants to create output that is creatively inferior to that created under lower constraints.

Finally, we find that the T80 treatment results in average image originality that is similar to the T10 treatment (2.81 vs 2.73, p>0.2). Taken together with our results on recognizability, our results show that participants in the T80 treatment submit images that are the same recognizability and originality as ones submitted in the T10 treatment, resulting in participant performance that is exceptionally poor. Our results suggest that setting high recognizability constraints is an ineffective way to encourage participants to factor in a sufficiently high level of recognizability in their creative output.

We now understand how image originality, image recognizability, and participant payoffs vary across treatments. Our goal now is to explore the behavioral mechanisms that could be driving our results. We start by comparing the originality-recognizability trade-off across treatments to see if and how they differ across our four treatments (a different originality-recognizability trade-off suggests that participants might be approaching the task differently). Then, we examine the survey responses of participants to see if and why participants in the T80 treatment are consciously ignoring or under-investing in the recognizability

Figure 3.4: Image Originality Versus Recognizability



(a) Average Originality Across Treatments

(b) Linear Fit Per Treatment

Notes: Panel (a) shows average image originality versus recognizability across treatments and Panel (b) shows a linear fit of average image originality versus recognizability in each treatment.

constraint.

3.4.5 Image Recognizability and Originality Across Treatments

Comparing the originality-recognizability trade-off across treatments can help us understand if participants are approaching the task similarly or different across our treatments. Figure 3.4(a) plots the average image originality versus recognizability across our experimental treatments. As implied previously, we find that images with higher recognizability generally have lower originality (nptrend test, p=0.00 for all treatments).

Our goal now is to see if the trade-off is the same across treatments. In Figure 3.4(b), we plot a linear fit of originality versus recognizability in each treatment. We notice that the trade-off between originality and recognizability is similar between the T0, T10, and T80 treatments, but that it is noticeably worse in the T40 treatment. To confirm this, we repeat our previous regression on image originality across treatments, controlling for image recognizability in Table 3.2 Column 2. We find that, when controlling for recognizability, the significance of the treatment dummies for the T10 and T80 treatments disappear, while the coefficient remains significant for the T40 treatment (p=0.00). This implies that the originality-recognizability trade-off remains the same in the T0, T10, and T80 treatment, but becomes significantly worse in the T40 treatment.

The similarity of the originality-recognizability trade-off between the T0 and T10 treatments can suggest that participants may be generating their creative ideas in a similar way.

Table 3.2: Regression of Image Originality

	No Controls	Accounting for Image Recognizability
Treatment 10	-0.332**	-0.219
	(0.136)	(0.134)
Treatment 40	-0.576***	-0.494***
	(0.132)	(0.130)
Treatment 80	-0.265*	-0.183
	(0.149)	(0.145)
Recognizability		-0.00827***
		(0.00103)
Constant	3.068***	3.361***
	(0.104)	(0.109)
Observations	1,864	1,864
R-squared	0.019	0.057

Robust standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

Despite the same trade-off between the two, we previously found that images in the T10 treatment are significantly more recognizable and more original than the T0 treatment. Unlike participants in the T0 treatment, who can submit any image and receive payment, participants in the T10 treatment must submit images that are at least minimally recognizable. Based on our results, we posit that participants in the T10 treatment might also be focusing on originality and are creating images as they would in the T0 treatment (resulting in the same originality-recognizability trade-off), but might be "filtering out" images they believe are not minimally recognizable (submitting those images that are more recognizable and less original).

The T40 treatment results in an originality-recognizability trade-off that is worse than the T0 and T10 treatments. This suggests that they are approaching the task differently than the T0 and T10 treatment, resulting in a creatively worse set of images. Unlike participants in the T10 treatment, participants in the T40 treatment must submit images that are fairly recognizable. We posit that, rather than focusing on originality goal and filtering out non-recognizable images, participants in the T40 treatment might be pursuing a goal to create images that are both recognizable and original. Pursuing both opposing goals could be why we see images that are creatively inferior to the T0 and T10 treatment.

Finally, our results show that the originality, recognizability, and originality-recognizability trade-off of images is the same in the T10 and T80 treatments. This suggests that participants in both treatments might be approaching the the creativity task in a similar way, placing a similar emphasis on originality and a similar emphasis on recognizability. While this strategy would fairly work for participants in the T10 treatment where the constraint is low, participants in the T80 treatment are substantially under-emphasizing the high constraint they must meet (resulting in their poor performance). We now look a participants' survey responses to see how participants in the T10 and T80 could be possibly approaching the creative task.

3.4.6 Survey Responses on Factoring Recognizability

Our first goal is to see how participants are approaching the task in the T10 treatment. Specifically, we want to see if they are consciously placing more emphasis on the recognizability constraint than in the T0 treatment. We test this by examining the participant responses to our open-ended survey questions and flagging the instances where they explicitly stated that they factored in recognizability in their images. We find that the percentage of participants explicitly stating that they factored in recognizability is higher in the T10 than the T0 treatment (32.7% vs 15.8%, p=0.06). This suggests that, as expected, participants in the T10 treatment seem to place a greater emphasis on recognizability in their responses than in the T0 treatment.

Our second goal is to how participants are approaching the creative task in the T80 treatment. Interestingly, we find that participants in the T80 treatment do not explicitly say they factor in the recognizability constraint more often than in the T10 treatment (32.7% vs 32.2%, p>=0.2). Our goal now is to see think that their images would automatically meet the 80% constraint (allowing them to approach the task similarly to the T10 treatment). In the exit survey, we ask participants to state the average recognizability of their submitted images. We find that participants in the T80 treatment believe their images will have a significantly higher recognizability than in the T10 treatment (69.2% vs 58.6%, p=0.00). This shows that they participants in the T80 treatment are significantly more overconfident in the recognizability of their submitted images compared to those in the T10 treatment (difference compared to actual: 23.6% vs 10.2%). Importantly, the average recognizability stated in the T80 treatment is still lower than the 80% required to receive payment. This suggests that participants in the T80 treatment may be consciously under-investing in the recognizability constraint, which we posit might be because of the task's perceived difficulty.

On a final note, participants, in our experiment, are not given any example images and do

not receive any feedback on the eight images they submit (a one-shot experiment). We now run further analysis on our survey responses to see how accurately participants can predict an image' recognizability and originality.

3.5 Participant Beliefs on Image Originality and Recognizability

Participants, in our experiment, do not receive any feedback on the recognizability and originality of their images before they submit them. We now check to see how accurately participants can predict the originality and recognizability ratings their images will receive and if prediction accuracy differs between high and low performers. In our exit survey, we ask participants to specify what they think will be the average recognizability and average originality ratings their images will receive. Our results show that participants overestimate both the recognizability and originality their images will receive, overestimating the recognizability of their images by 40% and the originality of their images by 25% on average. We also ask participants to specify which of their submitted images will receive the highest originality rating and the highest recognizability rating. Our results show that the image participants choose to be the most original, on average, ranks 3.5 in originality among their submitted images, while the image they choose to be the most recognizable, on average, ranks 4.3 in recognizability among their submitted images. We perform a median split on performance in each treatment and find that high performers in our experiment are significantly better able to predict which of their images will receive the highest originality rating compared to low performers (ranksum test, p=0.00), but not which one will receive the highest recognizability (p=0.17). Our results suggest that participants are better at predicting the originality rather than the recognizability of their images. A natural follow-up for our experiment could see how participant performance differs if participants perform the experiment across multiple rounds where they receive intermediate feedback.

3.6 Discussion and Conclusion

Knowledge workers are often asked to create output that is original and that meets a certain level of usefulness for the company. While the creativity literature suggests that it is difficult to create output that is both original and useful, it is not clear how varying constraints on the usefulness of a knowledge worker's output specifically affects the originality and usefulness of their work and their overall performance (Berg, 2014). We focus on settings where a likely

trade-off exists between the originality and the usefulness of a knowledge worker's creative output. Our goal is to see how a participant's performance on an originality-focused creative task varies with increasing recognizability constraints and if we can improve participant performance by "artificially" varying the recognizability constraint we set.

Our paper examines this setting by running a lab experiment that asks participants to work on a creative task with an originality goal and a usefulness constraint that we vary across four treatments. The creative task asks participants to create images using a set of building images and emoji. Participants are paid based on the originality of each submitted image on condition that it passes the usefulness constraint. We measure originality based on how different a submitted image is to other images in the experiment and we measure usefulness based on how recognizable the image is to a set of untrained raters. Our "baseline" T0 treatment sets a 0% recognizability constraint that incentivizes participants based on the originality of their images regardless of their recognizability. This is designed to encourage participants to focus only on originality. Our T10 treatment adds a 10% recognizability requirement that is designed to encourage participants to account for a minimal level of recognizability as their pursue an originality goal. We then set a 40% recognizability constraint in the T40 treatment that is designed to encourage participants to focus on both originality and recognizability in their creative output. Finally, our T80 treatment adds an 80% recognizability constraint designed to encourage participants to emphasize recognizability despite the stated originality goal of their task.

Our results show that participants in the T0 treatment, who are likely pursuing an originality-only goal for their images, generate images that are the most original and least recognizable compared to the other treatments, which results in participant payoffs that are the highest subject to the 0% constraint. We find that introducing a low 10% recognizability successfully encourages participants in the T10 treatment to factor in more recognizability in their creative output. This results in images that are significantly more recognizable and less original than the T0 treatment, improving participant performance subject to the 10% constraint. We posit that participants in the T10 treatment might be pursuing an originality goal for their images, as in the T0 treatment, but "filter out" images that they think are not minimally recognizable.

We find that setting moderate or high recognizability constraints results in surprising effects on a the recognizability, originality, and a participant's overall payoff. Our results show that setting a moderate 40% recognizability constraint in the T40 treatment results in relatively poor participant performance subject to the 40% recognizability constraint. We find that participant performance is surprisingly higher in the T10 treatment than the T40 treatment subject to the same 40% recognizability constraint. This is because the T10

treatment results in images with similar recognizability as in the T40 treatment, but with significantly higher originality. This shows that participants in the T40 treatment can benefit from pursuing an "artificially" lower constraint. We posit that, unlike participants in the T10 treatment, participants in the T40 treatment might be pursuing a goal to create images that are both an original and a recognizable, which results in images that are creatively inferior to the T10 treatment.

Finally, we find that the high 80% constraint in the T80 treatment results in participant performance that is exceptionally poor and that can be improved by having participants simply submit eight purely recognizable images. Our results show that participants in the T80 treatment submit images with the same originality and recognizability as in the T10 treatment. We posit that participants in the T80 treatment might be approaching the task as they would in the T10 treatment and that they are consciously under-investing in the high recognizability constraint.

Our paper shows that low usefulness requirements can be an effective way for managers to improve employee performance by encouraging them to factor in usefulness as they pursue an originality goal. We show that setting moderate usefulness requirements can be detrimental to employee performance by causing employees to generate output that is creatively inferior. Managers, in such cases, can improve employee performance by "artificially" lowering the usefulness constraint they set. Finally, we show that high usefulness constraints can be an ineffective way to encourage employees to factor in usefulness substantially in their creative output, resulting in output that rarely meets the usefulness constraint. Managers can, in such cases, improve employee performance by changing the goal they set for employees to emphasize usefulness rather than originality.

In our paper, we focus on settings where there is a clear trade-off between originality and usefulness. We measure originality based on how different a participant's creative output is from that of other participants and measuring usefulness based on how recognizable a participant's creative output is to a set of untrained raters. Future research should consider other measures of originality and usefulness and should consider settings where the trade-off between the two may not be as clear. Furthermore, we focus on knowledge worker settings where employees are mainly concerned with original ideas that may also need to pass a minimum level of usefulness. Future research can instead focus on work settings where employees are mainly concerned with the usefulness of their ideas that might need to pass a minimum level of originality. Participants in our experiment also do not receive any feedback for their images before they are submitted. Future research can focus on settings where participants perform a creativity task over multiple rounds and receive feedback between the rounds.

CHAPTER 4

Task Switching Behavior and Knowledge Worker Productivity

4.1 Introduction

Knowledge workers are often assigned creative tasks to work on that they must complete in a timely manner (Drucker, 1999; Hopp et al., 2009). For example, a research engineer working on a car's safety features can be asked to generate new simulations to test those features before it ships to customers. For that, she would need to think of scenarios that customers are likely to encounter and that the company has not tested yet, think of ways to realistically capture those scenarios in the simulation software, and to program and implement the simulation. Many knowledge workers can often find themselves "stuck" or unable to find a solution to a problem while working on their assigned creative tasks. Previous research has shown that setting a creative task aside to work on another task or to take a break can help a knowledge worker get "unstuck," possibly by giving them time to incubate on a solution and/or by forcing them to mentally set the task aside, allowing them to approach the problem with a fresh mindset when they return to it (Smith and Blankenship, 1991; Sio and Ormerod, 2009; Gilhooly, 2016).

In addition to their assigned creative tasks, knowledge workers can also have other tasks of different types that they need to work on. Unlike creative tasks, whose solution involves at least some level of ambiguity, these other tasks could be repetitive (and non-creative) in that they can often be completed by following a number of clear and precise steps. For example, in addition to her work on creating new simulations, the research engineer may also have to check the results of simulations submitted by her team members and flag any issues she finds. Depending on the setting, a knowledge worker may find herself forced to switch between two such tasks (Mortensen and Gardner, 2017). While the creativity literature suggests that switching can be beneficial for creative task performance, a separate literature has shown that task switching can be harmful to performance on repetitive tasks (Allport

et al., 1994; Monsell, 2003; Strobach et al., 2012). As a knowledge worker, it is then unclear how switching between both assigned tasks would affect one's performance on each.

In certain settings, a knowledge worker can be given some discretion in how they switch between their assigned tasks (Madjar and Shalley, 2008). Rather than switching in a predetermined way, the knowledge worker would then need to decide if and when she wants to switch tasks. Previous experimental literature has shown that participants given the discretion to switch between two creative tasks rarely do so and miss out on the performance benefits of task switching (Lu et al., 2017; Madjar and Shalley, 2008). It is unclear if knowledge workers would behave differently if the two tasks are of different types (repetitive and creative), with either one possibly serving as a mental break from the other. Furthermore, in certain knowledge worker settings, a manager may intervene by nudging their employee to switch tasks if they think it is in their interest to do so. For example, the research engineer's manager could suggest she take a break from working on a simulation if she has not made progress on it in a while. While such behavioral interventions exist in a variety of real-world knowledge worker settings, to the best of our knowledge, no research has been done to study their effect on a knowledge worker's discretionary switching behavior in a lab setting. As such, in our paper, we conduct a lab experiment to answer the following research questions: How does switching between a creative task and a repetitive task affect a knowledge worker's performance on each task and their overall performance? How do knowledge workers switch between tasks if they are given the discretion to do so? Can behavioral nudges be used to guide a knowledge worker's switching behavior and improve their performance?

Our experiment consists of five treatment that vary in a participant's ability and freedom to switch between a creative task and a repetitive task and in the presence of nudges encouraging them to switch tasks. In the Forced No Switch [FNS] treatment, participants are asked to work on one task (either creative or repetitive) before permanently switching to the other task (either repetitive or creative). In the Forced Switch [FS] treatment, participants are forced to repeatedly switch between the creative task and the repetitive task. We, then, run three discretionary switch treatments that allow participants to freely switch between the creative and repetitive tasks. The Discretionary Switch - No Nudge [DNN] treatment allows participants to switch, but does not give them any nudge to switch between treatments. Our final two treatments give participants different types of nudges, in the form of a pop-up message, to encourage them to switch. The Discretionary Switch - Time-Based Nudge [DTN] treatment gives participants a nudge if they haven't switch from either task in a while. Finally, the Discretionary Switch - Progress-Based Nudge [DPN] treatment nudges participants to switch tasks if they have not made progress on the creative task in a while or if they have not switched from the non-creative task in a while. Switching in the [DNN],

[DTN], and [DPN] treatments is completely voluntary and participants in the [DTN] and [DPN] treatments can simply "ignore" any nudges to switch that they receive.

Our results show that, as seen in the creativity literature, forcing participants to switch between tasks in the [FS] treatment substantially improves their performance on the creative task compared to participants in the [FNS] treatment, who are forced not to switch. This provides further evidence that setting a task aside to work on an unrelated task can help participants find solutions to creative questions. Similarly, as seen in the repetitive task switching literature, switching between the creative and repetitive task significantly lowers a participant's performance on the repetitive task. This results in overall performance that is not significantly higher in the [FS] treatment compared to the [FNS] treatment. We also find that participants, given the discretion to switch tasks in the [DNN] treatment, rarely switch between tasks. Interestingly, this significantly lowers their performance on the repetitive task without improving their performance on the creative task. This results in participant performance in the [DNN] treatment that is lower than the [FNS] treatment and significantly lower than the [FS] treatment. We show both time-based and progress-based behavioral nudges in the [DTN] and [DPN] treatment are surprisingly effective ways to encourage participants to voluntarily switch between tasks more often, resulting in task performance that is similar to the [FS] treatment.

Our paper highlights the benefits and the potential pitfalls of task switching when knowledge workers are assigned both creative and repetitive tasks. Importantly, we show that knowledge workers may not switch between tasks sufficiently if given the discretion to do, which can result in worse performance than their not switching tasks or repeatedly switching tasks. WE show that, in such cases, a manager can improve employee performance by using behavioral nudges to encourage their employees to voluntarily switch between tasks more often.

4.2 Literature Review

We now discuss how our paper contributes to the operations management literature studying knowledge worker productivity and to the literature studying task switching between creative tasks and between repetitive tasks. We then discuss how our nudge treatments build on insights from previous experimental work on reminders.

4.2.1 Knowledge Worker Productivity in Operations Management

Performance and productivity have been long-standing topics of interest in operations management (Smith and Robey, 1973; Ebert, 1976; Fujimoto and Clark, 1991; Herroelen and Leus, 2005; Schmenner, 2015). Much of the productivity research in operations has focused on improving productivity in settings traditionally associated with standardized work, where tasks are often physical and repetitive in nature, sometimes referred to as blue-collar work. This research considers topics such as the effects of work sharing, individual and group incentives, task switching, inventory policies, and queue structure on productivity (Schultz et al., 1999; Shunko et al., 2018; Stratman et al., 2004; Bendoly et al., 2014; de Vries et al., 2016). As opposed to production, service and professional jobs are typically more creative and knowledge intensive, are inherently less certain that physical tasks, and give employees greater discretion in carrying out tasks based on their judgment and learning capabilities (Spohrer and Maglio, 2008). While task switching has been repeatedly shown to lower performance in blue-collar settings, the creativity literature (discussed below) suggests that switching tasks can, in some cases, improve performance on a creative task. We explore this in our paper by directly studying the effect of switching between a creative and a repetitive task on a participant's performance. Furthermore, unlike most blue-collar settings, knowledge workers can in some cases be given some discretion in how they switch between their assigned tasks. We explore this setting in our paper by giving participants the freedom to freely switch between tasks.

4.2.2 Task Switching Between Repetitive Tasks

There is extensive experimental literature studying the effect of task switching on repetitive and non-creative tasks. The tasks considered in this literature are non-creative in that they have a clear solution that can always be achieved by following a clear and specific number of steps or actions and they are repetitive in that different questions differ in their prompt but can otherwise be solved following the same steps (Monsell, 2003). For example, in their seminal work, Allport et al. (1994) present participants with a number between 1 and 9 that is repeated between 1 and 9 times and ask them to decide, in one task, if the number presented is less than or greater than 5 and, in the other task, to decide if the number is repeated less than or more than 5 times. They show that participants that are forced to repeatedly switch between the two non-creative and repetitive tasks take substantially more time to answer each question and are more likely to answer a question incorrectly compared to participants that do not switch. This finding has been shown consistently in the task-switching literature across a broad range of repetitive tasks. We refer to the reader to Kiesel

et al. (2010) for an excellent review of the task switching literature.

In our paper, we consider situations where a knowledge worker must work on both a creative (and non-repetitive) task and a repetitive (and non-creative) task. The repetitive task we use is based on the work of Brüggen et al. (2018) and is in the form of a letter look-up task that asks participants to replace letters with numbers from a look-up table. Similar to the task-switching literature, we expect participants to incur a switching cost when they switch between tasks, which can decrease their performance on the repetitive task. While participants in the literature often switch tasks every few seconds, participants in our setting switch tasks every minute, which can result in a much lower switching cost. Furthermore, the look-up table participants use for decoding letters into numbers flips every question, forcing participants to incur a small setup cost when they go from one question to the next in the repetitive task, further diminishing the relative cost of switching away from the repetitive task. Overall, we expect that task switching will result in a decrease in performance on the repetitive task in our setting, but may result in an improvement in performance on the creative task which we discuss below.

4.2.3 Task Switching between Creative Tasks

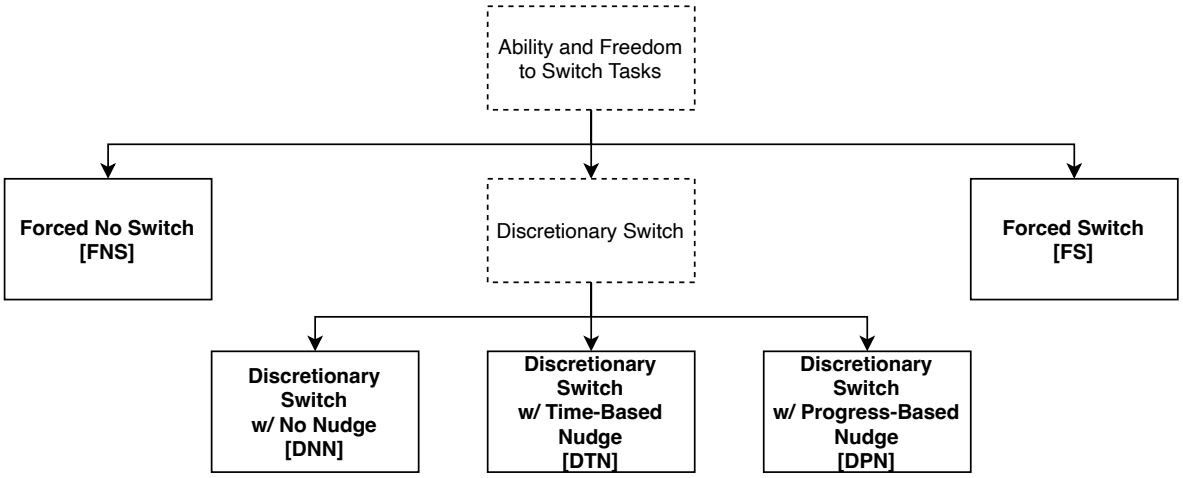
The creativity literature broadly categorizes creative thinking into convergent thinking, where the goal is often to find one correct solution to a clearly defined problem, and divergent thinking, where the goal is to generate a number of new, usually diverse, ideas in a context where more than one solution may exist (Colzato et al., 2012; Guilford, 1967; Duncker and Lees, 1945). While different knowledge workers may encounter either (or both) types of creative tasks in their work, in our paper, we focus on convergent creative tasks, where a knowledge worker has a clearly defined problem and is searching for one correct solution to that problem. One common way of capturing convergent thinking in an experimental setting, which we use in our paper, is through a Remote Associates Test or RAT (Mednick, 1962; Wu et al., 2020). An RAT asks participants to find one word that has a clear connection to each one of three, in many cases, unrelated, cue words. For example, a participant is presented with the words "stick/birthday/light" to which the solution is "candle" as in "candlestick," "birthday candle," and "candle light."

An emerging literature has shown that setting a creative task aside can improve performance by either making participants more original or helping them find a solution to a creative task (Sio and Ormerod, 2009). Participants can get "stuck" working on a RAT question if they fixate on a word that is clearly associated with one of the three cue words, but is unrelated to the other two words. An emerging literature has shown that setting aside

a convergent creative task and working on an unrelated task can increase the likelihood that a participant finds a solution Sio and Ormerod (2009). To test the benefit of task switching to overcome fixation, Smith and Blankenship (1991) asked participants to work on a RAT with a fourth misleading answer to push participants to fixate on an incorrect answer. They showed that participants that switched away from the RAT to an unrelated task (reading a science fiction novel for five minutes) were significantly more likely to find a solution to their assigned RAT questions compared to participants that were asked to continue working on the RAT without switching. The literature has shown that the type of interrupter task and the length of interruption is important to help participants find a correct solution. In our paper, we give a participants a repetitive non-creative task, in the form of a letter look-up task, that they can do while not working on the RAT. This is a task that participants must approach in a different way compared to an RAT, where they need to manually look up each letter for its corresponding code in a look-up table. We expect that switching between the RAT and the letter look-up task will increase participant performance on the RAT by helping participants get "unstuck" finding solutions to RAT questions.

Knowledge workers are often given the discretion to switch between their assigned tasks. Madjar and Shalley (2008) show that if given the discretion to switch between two divergent creative tasks and one repetitive task, participants rarely switch between tasks resulting in performance similar to those forced not to switch. Lu et al. (2017) similarly find that participants given the discretion to switch between two convergent creative questions, in the form of two RAT questions, perform similarly to those forced not to switch, whereas participants forced to switch perform significantly better. Our paper builds on the work of Lu et al. (2017) by also focusing on the effect of task switching on performance in on a convergent creative task. We expect that participants, given the discretion to switch tasks, will also rarely switch between their assigned tasks in our setting. Unlike Lu et al. (2017), we focus on a setting where knowledge workers (and participants) must work on two different types of tasks, one convergent creative task and one repetitive non-creative task, and where performance (and participant payoff) depends on their performance on both tasks (rather than participants receiving a flat rate). Previous research has shown that "reminding" individuals to perform a task might be enough for them to do so (Calzolari and Nardotto, 2017). In our paper, we study the effect of different nudges to encourage participants to voluntarily switch between tasks more often and the effect that has on their performance on both the creative and repetitive tasks.

Figure 4.1: Experimental Treatments.



4.3 Experimental Design

Our experiment consists of three stages. In Stage 1, participants spend 2.5 minutes working on a non-creative repetitive task consisting of 15 questions and, in Stage 2, they spend 2.5 minutes working on a creative non-repetitive task consisting of 10 creative questions. They then spend 30 minutes in Stage 3 working on one non-creative task consisting of 90 questions and one creative task consisting of 60 questions, each for 15 minutes. Our treatments then vary a participant's ability and freedom to switch between the creative and non-creative tasks in Stage 3. At the end of the experiment, participants are paid based on the percentage of non-creative questions they solved in Stage 1, creative questions they solved in Stage 2, non-creative questions they solved in Stage 3, and the creative questions they solved in Stage 3.

The non-creative task is a letter look-up task that asks participants to replace each letter in a four letter code with its corresponding numbers from a look-up table. For example, a letter code such as "A-V-X-S" would be "151-499-811-288" based on a given look-up table. There are two look-up tables, one for even numbered questions and one for odd numbered questions and each letter has a three digit number associated with it. This makes it difficult for participants to remember a three digit code without looking at the table, forcing participants to continuously look at the table and come back. This makes it less likely for a participant to become "more efficient" at solving a letter look-up task the longer they work on it. The creative task is a remote associations task or RAT (Bowden and Jung-Beeman, 2003). Each RAT question consists of three words and the participant's task is

to find a fourth word that has a clear connection to each of the three words. For example, for the three prompt words "stick/birthday/light" the answer is "candle." In each task, a participant must answer a question correctly before they can move to another question. A participant can "skip" a question after spending one-fifth the task time working on it (they can skip after 30 seconds in Stage 1 and Stage 2 and after 180 seconds in Stage 3). This makes it so that participants are not able to switch between questions in the same task and can only switch between tasks (in the treatments that they are allowed to switch), which allows us to measure the effect of task switching more directly.

4.3.1 Participants

Two hundred and fifty participants are recruited at the University of Michigan. The experiment was conducted online using zTree Unleashed and Zoom. Participants are paid a $5 show-up fee and based on their performance in the experiment. Average participant payment was $17 and payments ranged between $10 and $25.

4.3.2 Procedure

Participants are sent a Zoom invite and, upon joining the Zoom session, are privately sent an individualized link to Ztree Unleashed. Participants are asked to have their video camera on for the duration of the experiment and are not allowed to talk with one another. At the start of each stage in the experiment, the instructions are read aloud. Participants are then asked to work on a set of practice questions to make sure they understand how to solve each assigned task and how their payoff is calculated. We then go through the practice questions together and answer any questions they have. At the end of the experiment, participants are informed of their payoff on each stage and their final payoff.

4.3.3 Treatments

Our experiment consists of five treatments. In the Forced No Switch [FNS] treatment, participants randomly spend 15 minutes working one task (either the creative or non-creative task) before permanently switching to the other task. This treatment allows us to measure baseline performance, i.e. how many creative and non-creative questions a participant can solve correctly if they are not allowed to switch between tasks. In the Forced Switch [FS] treatment, participants are forced to switch between tasks every minute. Upon switching between tasks, their progress is saved and they can continue working on the task when they switch back. This treatment allows us to see if a participant benefits from switching between

a creative and non-creative task.

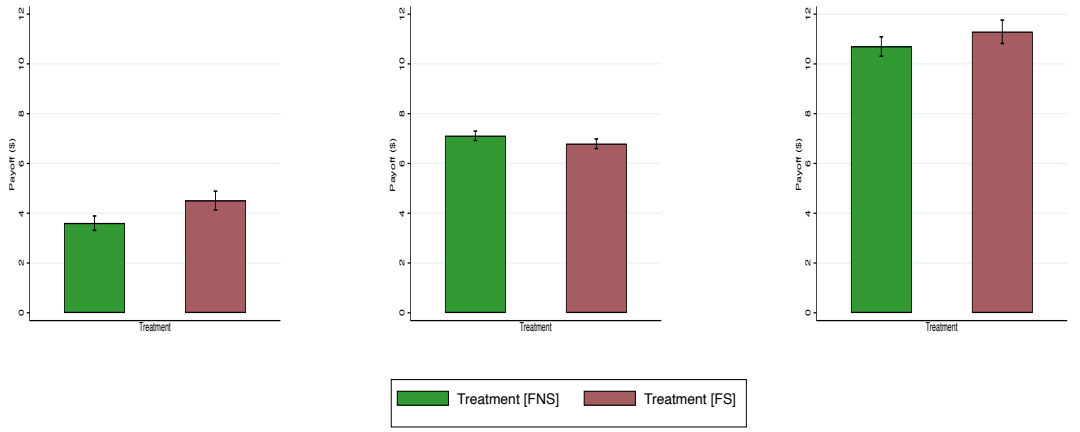
We then run three discretionary switch treatments that allow participants to switch between tasks whenever they want by pressing a "Switch Tasks" button. The treatments differ in a presence of a "nudge" that serves as a suggestion to switch between tasks and the type of that nudge. The Discretionary Switch with No Nudge [DNN] treatment allows participants to switch between tasks but does not have any nudge for them to switch. This treatment allows us to measure a participant's tendency to voluntarily switch between tasks. The Discretionary Switch with a Time-Based Nudge [DTN] treatment nudges participants to switch between tasks if they have spent the last 60 seconds working on either the creative or the non-creative task without switching. The nudge is in the form of a pop-up that automatically appears on their computer screen. Participants must press "Okay" to acknowledge they saw the nudge, at which point, they can choose to continue working on their task or to switch to the other task. If a participant spends another 60 seconds working on the same task without switching, they will receive another nudge. This treatment is similar to the [FS] treatment, but gives participants the freedom to not switch after 60 seconds if they want to.

Finally, the Discretionary Switch with a Progress-Based Nudge [DPN] treatment includes two types of nudges. Similar to the [DTN] treatment, a participant is nudged if they have been working on the non-creative task for the last 60 seconds without switching. A participant is nudged in the creative task if they have been working on it continuously and have not solved a question correctly in a while. Specifically, twice the amount of time it took a participant to answer a question correctly in Stage 2. Receiving a nudge in the creative task is a signal that a participant is likely "stuck" working on a creative question and should consider switching to the other task. This treatment allows us to see if participants benefit from receiving a nudge when they are likely stuck working on a creative question.

4.4 Results

We start by studying the effect of frequent forced task switching on a participant's performance on a creative task and a repetitive task and on their overall performance. We then explore how participants switch between tasks if given the discretion to do and the effect that has on their performance. Finally, we see if certain behavioral interventions, in the form of nudges to switch, can be used to guide a participant's switching behavior and improve their performance. Our results show that task switching has a significant, but opposite effect on a participant's performance on each task, resulting in similar overall performance. Interestingly, we find that discretionary task switching results in, sometimes significantly, worse performance than either not switching or frequent forced switching. Our paper finds that

Figure 4.2: Participant Payoffs in the Forced No Switch and Forced Switch Treatments



(a) Creative Task (b) Repetetive Task (c) Both Tasks

Note: The above graphs shows participant payoffs in the [FNS] and [FS] treatments on the creative task, repetitive task, and on both tasks together.

either performance-based or time-based nudges are a surprisingly effective way to improve participant performance by encouraging participants to voluntarily switch more often.

4.4.1 Forced Task Switching

To study the effect of frequent forced task switching, we compare participant payoffs in the [FNS] treatment, where they are forced not to switch tasks, to to [FS] treatment, where they are forced to switch tasks every minute. Figure 4.2 displays participant payoffs in the creative task and the repetitive task and on both tasks for the [FS] and the [FNS] treatments. As a reminder, participants are incentivized based on the number of creative questions and repetitive questions they complete. The creative task is in the form of an RAT and the repetitive task is a letter look-up task.

4.4.1.1 Creative Task Payoff

Participants working on the creative task might get "stuck" finding a solution to a creative question, which can lower their payoff by leaving them with less time to work on the subsequent creative questions. Forcing a participant to switch to the repetitive task can improve a participant's performance on the creative task by helping them get "unstuck" finding a solution. It can also harm a participant's performance on the creative task by forcing them to pay a mental setup cost every time they switch tasks and by forcing them to switch even

when they are not "stuck" on a creative question. Interestingly, in Figure 4.2(a), we show that the [FS] treatment does significantly improve participant payoff on the creative task by 25% over the [FNS] treatment (we discuss treatment significance in our regression analysis in the section below). This shows that despite the additional mental setup cost, forcing participants to switch does improve their performance on the creative task, likely by helping them get unstuck finding solutions to their creative questions.

4.4.1.2 Repetitive Task Payoff

Participants working on the repetitive task must follow the same number of steps to find the answer to each repetitive question. Forcing a participant to switch to the creative task can lower a participant's performance on the repetitive task by forcing them to pay a mental setup cost. While the steps are the same across repetitive questions, the questions are different and force participants to pay a small mental setup cost between questions. As such, switching to the creative task may not substantially lower a participant's performance. In Figure 4.2(b), we show that the [FS] treatment does significantly lower participant performance on the repetitive task by 4.5% over the [FNS] treatment. This shows that switching to a creative task does lower a participant's performance on the repetitive task even when a participant has to pay mental setup costs across repetitive questions.

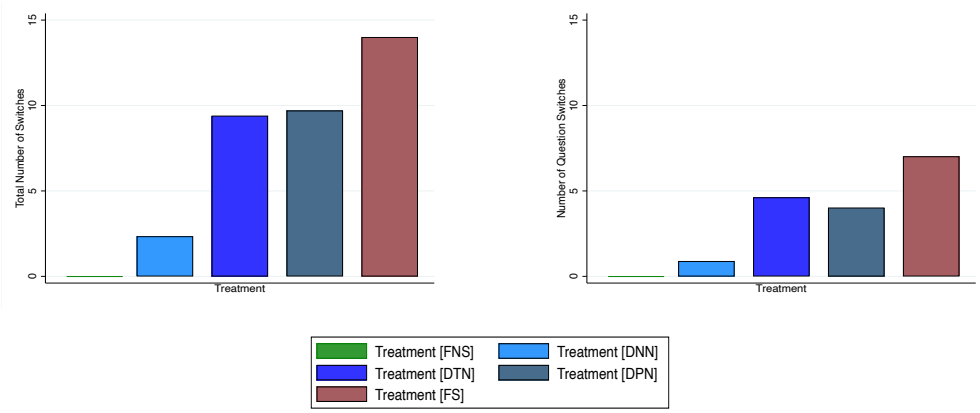
4.4.1.3 Overall Payoff

Our results show that task switching has significant, but opposite effects on a participant's payoff on the creative task and the repetitive task. Because participants in our setting are incentivized equally on both tasks and they complete more of the repetitive task, the repetitive task accounts for a larger percentage of their overall payoff. As such, in Figure 4.2(c), we find that switching between tasks only results in a non-significant 5.4% improvement in a participant's overall payoff. Our results highlight the benefit of task switching on creative task performance but caution against their potential drawbacks on a repetitive task especially when participants are incentivized equally for both.

4.4.2 Discretionary Task Switching

Participants in the [FS] treatment must switch between tasks every minute, which likely forces them to switch even when they are not stuck, forces them to wait even when they are stuck, and possibly forces them to pay a needlessly high mental switch cost. Our goal now is to see how participants perform if they are given the discretion to switch. Participants with the freedom to switch can switch exactly when they need to, improving their performance

Figure 4.3: Switching Behavior Across Treatments



(a) Number of Switches Between Tasks (b) Number of Question Switches

Note: The above graphs shows switching behavior across the five treatments. In Panel (a) we show the total number of times a participant switched between treatments and in Panel (b) we show the number of questions they switched on.

on the creative task and reducing the effect of mental switching costs on their repetitive task performance. We directly test the effect of discretionary switching in the [DNN] treatment. We also check to see if we can improve a participant's discretionary switching behavior and performance with time-based nudges in the [DTN] treatment and with progress-based and time-based nudges in the [DPN] treatment.

4.4.2.1 Switching Behavior

We have shown that switching away from a creative question can help a participant find a solution to that question. Figure 4.3 displays the total number of times a participant switches from a task (ignoring when they switch when time runs out) and the number of creative questions a participant switches on across treatments. Interestingly, compared to the [FS] treatment, participants in the [DNN] treatment rarely switch tasks and switch on less than one creative question on average (2.3 switches vs 14 switch, p=0.00 and 0.89 questions vs 8.02 questions p=0.00). Given their poor switching behavior, it is unlikely that participants in the [DNN] treatment will see an improvement to their creative task performance. Surprisingly, we find that the [DTN] and [DPN] treatments are similarly effective at improving a participant's discretionary switching behavior over the [DNN] treatment (9.4 switches and 10.7 switches vs 2.3 switches, p=0.00 for both). Participants in both treatments also switch on significantly

Table 4.1: Regression for Creative Task Payoff

	No Controls	Accounting for Individual Ability	Accounting for Individual Ability and Task Switching
Treatment [DNN]	0.224	0.131	-2.615
	(0.426)	(2.616)	(2.524)
Treatment [DTN]	0.861*	5.760**	-3.559
	(0.444)	(2.710)	(3.079)
Treatment [DPN]	1.032**	5.538**	-2.628
	(0.439)	(2.708)	(2.965)
Treatment [FS]	1.185***	6.408**	-7.358**
	(0.449)	(2.751)	(3.623)
Individual Ability C		2.196***	1.959***
		(0.389)	(0.371)
Individual Ability R		0.806***	0.685**
		(0.286)	(0.271)
Number of Question Switches			1.727***
			(0.316)
Constant	3.502***	3.038	5.186
	(0.301)	(3.644)	(3.469)
Observations	251	251	251
R-squared	0.043	0.203	0.290

Standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

Note: Participants, across all treatments, spend Stage 1 and Stage 2 working on a non-creative repetitive task and a creative non-repetitive task, allowing us to control for their individual ability for creative tasks (Individual Ability C) and repetitive tasks (Individual Ability R) when studying the effect of task switching on performance.

more creative questions (4.6 and 4.0 vs 0.89, p=0.00 for both).

4.4.2.2 Creative Task Payoff

Our goal now is to see the effect of discretionary switching on participant performance on the creative task compared to the [FNS] and [FS] treatments. Our results show that the [DNN] treatment only results in a 3% increase in creative task performance over the [FNS] treatment, while the [DTN] and [DNN] treatments result in a 21.1% and a 23.2% increase respectively. To obtain more accurate treatment comparisons, we regress a participant's payoff on the creative task on treatment indicators in Table 4.1 Column 1 and accounting for differences

Table 4.2: Regression for Repetitive Task Payoff

	No Controls	Accounting for Individual Ability	Accounting for Individual Ability and Task Switching
Treatment [DNN]	-0.418	-0.429**	-0.465**
	(0.276)	(0.204)	(0.208)
Treatment [DTN]	-0.401	-0.432**	-0.554**
	(0.287)	(0.211)	(0.254)
Treatment [DPN]	-0.962***	-0.859***	-0.965***
	(0.284)	(0.211)	(0.244)
Treatment [FS]	-0.304	-0.418*	-0.597**
	(0.291)	(0.214)	(0.299)
Individual Ability C		0.0804***	0.0773**
		(0.0303)	(0.0306)
Individual Ability R		0.298***	0.296***
		(0.0223)	(0.0223)
Number of Question Switches			0.0226
			(0.0261)
Constant	7.093***	3.748***	3.776***
	(0.195)	(0.284)	(0.286)
Observations	251	251	251
R-squared	0.046	0.489	0.491

Standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

Note: Participants, across all treatments, spend Stage 1 and Stage 2 working on a non-creative repetitive task and a creative non-repetitive task, allowing us to control for their individual ability for creative tasks (Individual Ability C) and repetitive tasks (Individual Ability R) when studying the effect of task switching on performance.

in individual ability in Column 2. We measure individual ability as a participant's payoff on the creative task and repetitive task in the two practice stages. As expected, the [DNN] treatment results in a similar performance to the [FNS] switch and that is significantly worse than the [FS] treatment (post regression test, p=0.02). The [DTN] and [DPN] treatments result in a significant improvement in creative task performance that is similar to the [FS] treatment (post regression test, p>=0.02 for both). Table 4.1 Column 3, we find that the significance of the [DTN] and [DPN] treatments disappears when we account for differences in the number of creative questions switched, implying that the effectiveness of both treatments is due to changes in their task switching behavior over the [FNS] treatment.

Table 4.3: Regression on Overall Payoff

	No Controls	Accounting for Individual Ability	Accounting for Individual Ability and Task Switching
Treatment [DNN]	-0.195	-0.409	-0.840**
	(0.560)	(0.452)	(0.424)
Treatment [DTN]	0.460	0.432	-0.395
	(0.584)	(0.468)	(0.540)
Treatment [DPN]	0.0697	-0.0280	-0.500
	(0.577)	(0.468)	(0.535)
Treatment [FS]	0.881	0.544	-0.612
	(0.591)	(0.476)	(0.657)
Individual Ability C		0.410***	0.373***
		(0.0673)	(0.0622)
Individual Ability R		0.419***	0.371***
		(0.0494)	(0.0459)
Number of Creative			0.428***
Question Switches			(0.0627)
Number of Repetitive			-0.182***
Question Switches			(0.0414)
Constant	10.59***	4.204***	4.919***
	(0.396)	(0.630)	(0.588)
Observations	251	251	251
R-squared	0.016	0.372	0.475

Standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

Note: Participants, across all treatments, spend Stage 1 and Stage 2 working on a non-creative repetitive task and a creative non-repetitive task, allowing us to control for their individual ability for creative tasks (Individual Ability C) and repetitive tasks (Individual Ability R) when studying the effect of task switching on performance.

4.4.2.3 Repetitive Task Payoff

We now check to see if the effect of discretionary switching on participant performance in the repetitive task. We repeat our previous regression for repetitive task payoffs in Table 4.2. Surprisingly, we find that the [DNN] treatment results in repetitive task performance that is significantly worse than the [FNS] treatment, despite participants' rare switching. Both the [DPN] and [DTN] treatments also result in significantly worse performance on the repetetive task, similar to the [FS] treatment.

4.4.2.4 Overall Payoff

Finally, we check to see the effect of discretionary switching on a participant's overall payoff. We repeat our regression for total payoffs in Table 4.3. We find that all treatments result in overall payoffs that are similar to the [FNS] treatment. Importantly, the [DNN] treatment results in an overall payoff that is worse than the [FNS] treatment and significantly worse than the [FS] treatment (post regression test, p=0.04), while the [DPN] and [DTN] treatment result in similar payoffs (p>0.2). Our results show that participants, given the discretion to switch, rarely switch between tasks, which significantly decreases their performance on the repetitive task without improving their performance on the creative task. We find that setting time-based or progress-based nudges are surprisingly effective ways to improve participant performance by encouraging participants to voluntarily switch more often.

4.5 Discussion and Conclusion

Knowledge workers are often assigned both creative and repetitive tasks that they might be forced to switch between. Two streams of literature have separately examined the positive effect of forced task switching on creative tasks and the negative effect of forced task switching on repetitive tasks (Allport et al., 1994; Bowden and Jung-Beeman, 2003; Smith and Blankenship, 1991; Smith et al., 2017). Our first goal is to see the effect of forced task switching on a knowledge worker's performance on a creative task and a repetitive task. In some cases, a knowledge worker may also be given some discretion in how they switch between their assigned creative and repetitive tasks. Previous research has shown that participants, given the discretion to switch between two creative tasks, rarely switch tasks and do not gain any performance benefit (Lu et al., 2017; Madjar and Shalley, 2008). Our second goal is to see how participants, given the discretion to switch, would switch between tasks of different types, where one task can offer them a mental break from the other task. Finally, in a variety of knowledge worker settings, managers may nudge their employees to voluntarily switch between their assigned tasks. To the best of our knowledge, this has not been studied in an experimental setting before. Our third goal is to then study the effect of behavioral nudges on a knowledge worker's discretionary switching behavior.

Our paper runs a lab experiment that incentivizes participants based on their performance on a creative task and on a repetitive task. Our treatments then vary a participant's ability and freedom to switch between tasks and studies the effect that has on their task performance. We run five treatments. The [FNS] treatment forces participants not to switch tasks and the [FS] treatment forces them to switch tasks every minute. The [DNN] treatment then

gives participants the discretion to switch tasks. We then add time-based behavioral nudges for participants to voluntarily switch tasks in the [DTN] treatment and both time-based and progress-based behavioral nudges in the [DPN] treatment.

Our results show that, as seen in the creativity literature, forcing participants to switch tasks in the [FS] treatment substantially improves a participant's performance on the creative task compared to participants forced not to switch in the [FNS] treatment. This provides further evidence that setting a task aside to work on a separate, in this case repetitive, task can help participants find solutions to creative questions. We also find that, as seen in the repetitive task switching literature, switching between from a repetitive task to a creative task significantly decreases a participant's performance on the repetitive task. Taken together, our results show that switching between a creative and a repetitive task results in overall performance that is not significantly higher in the [FS] treatment compared to the [FNS] treatment. We also find that participants given the discretion to switch in the [DNN] treatment rarely switch between their assigned tasks. Interestingly, our results show that this significantly lowers their performance on the repetitive task compared to the [FNS] treatment without improving their performance on the creative task. This results in overall participant performance in the [DNN] treatment that is lower than the [FNS] treatment and significantly lower than the [FS] treatment. Finally, we find that both the [DTN] and the [DPN] treatments are surprisingly effective at improving participant performance by encouraging participants to voluntarily switch between asks more often.

Our paper suggests some potential benefits and drawbacks when knowledge workers are forced to switch between a creative and a repetitive task. We also show that knowledge workers given the discretion to switch tasks may not do so often, which can lower their performance compared to those asked not to switch or to switch often. We find that, in such cases, behavioral nudges can be an effective way for managers improve employee performance by encouraging them to voluntarily switch between tasks more often.

Our work speaks to situations where a knowledge worker's performance is judged equally based on their performance on tasks of different types. This could be in knowledge worker settings where both the creative task and the repetitive task are both integral to a project's overall progress and where progress on both is judged equally. In our future work, we want to also explore settings where the creative task is incentivized more heavily compared to the repetitive task. Based on the results of our paper, we believe that task switching, in such a setting, can improve a participant's overall payoff. In our paper, we focus on a convergent creative task and a repetitive task. This speaks to situations where a knowledge worker is working on a creative task with one clear solution that they might have trouble finding. Future research should explore the effect of task switching between a divergent creative task

and a repetitive task. That could speak to real-world settings where a knowledge worker's task is to generate new ideas or where the task has multiple solutions that may not be clear beforehand.

CHAPTER 5

Conclusion

The relatively complex settings that knowledge workers operate in provides a rich tapestry of problems that behavioral researchers can work on. In my dissertation, I investigate how a knowledge worker's productivity is affected by the design of the communication technologies they have access to, the usefulness constraints they must meet when pursuing a originality-focused creative task, and by their ability and freedom to switch between their assigned tasks. I hope that future behavioral research can build on my work on ESMPs to study other ESMP design features, the design of other communication technologies, and, more broadly, the design of other technologies that knowledge workers interact with. Separately, I hope that future research can also build on my work on creative tasks to study other aspects of creativity in knowledge worker settings, such as the effect of managerial feedback on creative output and the effect of creative task sequencing.

APPENDIX A

Appendix for "Enterprise Social Media Platforms and Knowledge Worker Productivity"

A.1 Panel Regressions on Helping Behavior

To complement our analysis of helping behavior in Section 2.4.2, we use panel regressions to compare helping behavior between treatments. Table A.1 shows the results from Probit, OLS, and Tobit panel regressions. The dependent variable, whether any help is given (Column 1) or the average help given (Columns 2 and 3) in a round of the social media stage, is regressed on treatment and round dummies. Many participants do not give any help through the platform, with some variation by treatment. Column 1 reports the results of a Probit regression where the dependent variable is an indicator which takes the value 1 if participant gave any help during the social media stage, and 0 otherwise. Here we see that the [B] treatment significantly increases the chance that a participant ever helps, while the [AI] and [FI] treatments do not significantly change the chance.

We next look at the amount of help given. Column 2 of Table A.1 reports the results of an OLS regression of the number of times an individual helps on treatment dummies. Because some of participants provide no help (i.e. the minimum value of help of zero), we also run a Tobit regression which accounts for this constraint. The results are reported in Column 3. Both analyses provide similar results: the [AI] treatment decreases the overall amount of help given, while the [FI] and [B] treatments increase the amount of help, with the badges treatment having a much larger positive effect on help.

For Study 2 treatments, the average amount of help given and requested across treatments in shown in Table A.2. As in the previous section, we compare help in the goal-only and badges treatments to the $[\overline{\emptyset}]$ treatment using Probit, OLS, and Tobit regressions. We find that all additional treatments significantly improve help over the $[\overline{\emptyset}]$ treatment.

Table A.1: Helping Behavior Across Lab Treatments

Method	Probit	OLS	Tobit
Dependent Variable	Indicator Help > 0	Amount Help Given	Amount Help Given
AI Treatment	-0.161	-0.611	-1.283***
	(0.114)	(0.435)	(0.490)
FI Treatment	0.124	0.148	0.535
	(0.0964)	(0.399)	(0.442)
B Treatment	0.269***	1.594***	2.487***
	(0.0843)	(0.383)	(0.472)
Individual Ability	0.00415	0.0676***	0.132***
	(0.00356)	(0.0194)	(0.0313)
Round 5	0.0568*	0.360***	0.543***
	(0.0344)	(0.100)	(0.201)
Round 6	0.00758	0.462***	0.606***
	(0.0368)	(0.144)	(0.201)
Round 7	0.0833**	1.045***	1.471***
	(0.0367)	(0.212)	(0.198)
Round 8	-0.0682**	0.197	0.0895
	(0.0323)	(0.170)	(0.205)
Constant	0.507***	-0.111	-2.630***
	(0.0868)	(0.409)	(0.656)
Observations	1,320	1,320	1,320
Number of Subjects	264	264	264

Standard errors in parentheses (clustered for Probit and OLS)
*** p<0.01, ** p<0.05, * p<0.1

Notes: Coefficients of Probit, OLS, and Tobit panel regressions are reported. The dependent variable in the Probit is an indicator variable for helping ($1_{help>0}$). The dependent variable in the OLS and Tobit is the total help given by a participant in a round. The lower level in the Tobit regression is specified at 0. Tobit regression results are similar when bootstrapped standard errors are used. Dummy variables are used to indicate the treatment and round number in the social media stage.

Table A.2: Helping Behavior Across Online Treatments

Method	Probit	OLS	Tobit
Dependent Variable	Indicator Help > 0	Amount Help Given	Amount Help Given
$\overline{G3}$ Treatment	0.213***	0.677**	1.499***
	(0.0801)	(0.287)	(0.407)
$\overline{B3}$ Treatment	0.257***	1.134***	2.016***
	(0.0913)	(0.429)	(0.406)
$\overline{BE}$ Treatment	0.208**	0.881**	1.623***
	(0.0878)	(0.425)	(0.388)
$\overline{BE5}$ Treatment	0.258***	0.882**	1.668***
	(0.0859)	(0.378)	(0.380)
$\overline{BE35}$ Treatment	0.244***	1.184***	2.038***
	(0.0796)	(0.361)	(0.397)
Individual Ability	0.00722***	0.0321**	0.0600***
	(0.00264)	(0.0126)	(0.0200)
Round 5	0.0652***	0.686***	1.013***
	(0.0237)	(0.0805)	(0.137)
Round 6	0.0628***	0.700***	0.981***
	(0.0237)	(0.0956)	(0.137)
Round 7	0.0628**	0.903***	1.272***
	(0.0261)	(0.115)	(0.136)
Round 8	0.00242	0.560***	0.803***
	(0.0275)	(0.111)	(0.137)
Constant	0.360***	0.00197	-2.072***
	(0.0915)	(0.309)	(0.468)
Observations	2,070	2,070	2,070
Number of Participants	414	414	414

Robust standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

Notes: Coefficients of Probit, OLS, and Tobit panel regressions are reported. The dependent variable in the Probit is an indicator variable for helping ($1_{help>0}$). The dependent variable in the OLS and Tobit is the total help given by a participant in a round. The lower level in the Tobit regression is specified at 0. Tobit regression results are similar when bootstrapped standard errors are used. Dummy variables are used to indicate the treatment and round number in the social media stage.

Table A.3: Canceling Statistics

Treatment	% Requests Canceled	Time to Cancel (sec)	% Request Again	Time to Request Again (sec)
$\emptyset_o$	14	48	83	55
G3	5	45	83	26
B3	9	42	82	43
BE	10	57	84	21
BE5	7	68	88	44
BE35	7	52	87	30

A.2 Differences in Request Canceling Behavior

Participants in our experiment can cancel their outstanding request at any time. They can do this if they a) answered the question themselves and have no need of the request anymore b) want to ask about a different question or c) waited too long to receive an answer and do not want to request help anymore. In Table A.3, we show the percentage of requests canceled, time to cancel, probability a participant requests help again in that round, and the time to request help again. We find that cancelling behavior, specifically, the time to cancel a request and the probability that a participant makes another request after cancelling are very similar across the basic, goal-only, and badges treatment. The percentage of requests cancelled and the time to request again after cancelling a request are slightly higher in the $[\overline{\emptyset}]$ treatment than the other treatments, but both are lower in the $[\overline{G3}]$ compared to the $[\overline{B3}]$ despite it having a marginally lower amount of helping behavior. We conclude that changes in canceling behavior are unlikely to be the major drivers behind the effectiveness of the badges treatments.

A.3 The Relationship Between Participant Performance and Requesting/Giving Help

We study the relationship between a participant's performance and helping behavior (help requested/given). Participants know how well they are performing in a given round, which can affect how much time they allocate towards helping others. A participant easily answering many questions may become satisfied with her progress, which may make her more likely

Table A.4: Help Given Based on Individual Ability

Treatment	**Location**	**Help Given By Low Individual Ability Participants**	**Help Given By High Individual Ability Participants**	**p-value (ranksum test)**
Ø	Lab	1	1.92	0.01
AI	Lab	0.49	1.34	0.01
FI	Lab	1.67	1.79	>0.2
B	Lab	2.82	3.53	0.12
$Ø_o$	Online	0.84	1.48	0.03
G3	Online	1.72	1.81	>0.2
B3	Online	2.13	2.43	>0.2
BE	Online	1.73	2.42	0.08
BE5	Online	1.86	2.3	>0.2
BE35	Online	2.44	2.17	0.19

to provide help than a person who is struggling through the questions. On the other hand, being quick to answer may make her more sensitive to the cost of two seconds she incurs for helping others. To study this relationship, we examine if having high individual ability is correlated with being more helpful and if being "unhelpful" is correlated with having low individual ability in the experiment. We do this by performing a median-split in each treatment based on individual ability and comparing the help given and help requested by high and low individual ability participants. Our results are shown in Table A.4 for help given and Table A.5 for help requested. We generally find that help given by higher individual ability participants is higher than that of lower individual ability participants, but the difference is only significant in the basic platform treatments and the [AI] treatment. Pooling all lab treatments together we get a rank sum test p-value of 0.00. This is in line with our OLS regression results in Table A.1 and Table A.2 that show that individual ability significantly increases the amount of help given. Our results also show a pattern of participants with higher individual ability requesting slightly more help than lower individual ability participants, but the difference is not significant in almost all treatments and not significant for all online or lab treatments pooled together (rank sum test min p> 0.2).

Table A.5: Help Requested Based on Individual Ability

Treatment	Location	Help Requested By Low Individual Ability Participants)	Help Requested By High Individual Ability Participants	p-value (ranksum test)
Ø	Lab	1.64	1.43	>0.2
AI	Lab	0.64	1.32	0.01
FI	Lab	1.91	1.69	>0.2
B	Lab	3.15	3.35	>0.2
$Ø_o$	Online	1.14	1.39	0.11
G3	Online	1.79	2.28	0.19
B3	Online	2.03	3.02	0.03
BE	Online	2.09	2.39	>0.2
BE5	Online	2.29	2.3	>0.2
BE35	Online	2.42	2.52	>0.2

A.4 Effect of Receiving Help

We would like to see how receiving help in one round affects your helping behavior in the following round. Specifically, we want to see if receiving help in one round encourages a participant to request more help in the following round and give help in the following round. In Table A.6, we run a panel regression on the help requested in one round on a participant's individual ability and the amount of help received in the previous round. We find that in all treatments participants who receive help are more likely to request help in the following round. We run the same panel regression in Table A.7 on the amount of help given. We find that receiving help in one round does not significantly increase the amount of help she gives in the following round in the [Ø], [AI], or [B] treatments. Interestingly, we find that receiving help in one round significantly increases help given in the following round in the [FI] treatment. We posit that the increased observability in the [FI] treatment, where other participants are able to see a participant's difference between help given and received, is likely to drive this significance. We also find that participants who request or give help in one round are very likely to request help and give help in the following round (with probability 0.87 and 0.74 respectively), while participants who do not request and give help

Table A.6: Help Requested Based on Previous Help Received

VARIABLES	Basic Treatment	AI Treatment	FI Treatment	B Treatment
Individual Ability	-0.0146	0.0318*	-0.0105	-2.83e-05
	(0.0169)	(0.0166)	(0.0149)	(0.0214)
Previous Help Received	0.664***	0.627***	0.507***	0.539***
	(0.0725)	(0.0533)	(0.0557)	(0.0647)
Constant	0.958**	-0.222	1.221***	1.667***
	(0.451)	(0.237)	(0.305)	(0.466)
Observations	240	240	336	240
Number of Subjects	60	60	84	60

Robust standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

in one round are also very likely to not request or give help in the following round (with probability 0.87 and 0.72 respectively). This suggests a partition in our experiment between participants that routinely request and/or give help and participants that routinely refrain from requesting and/or giving help.

A.5 Helping Trends Across Treatments

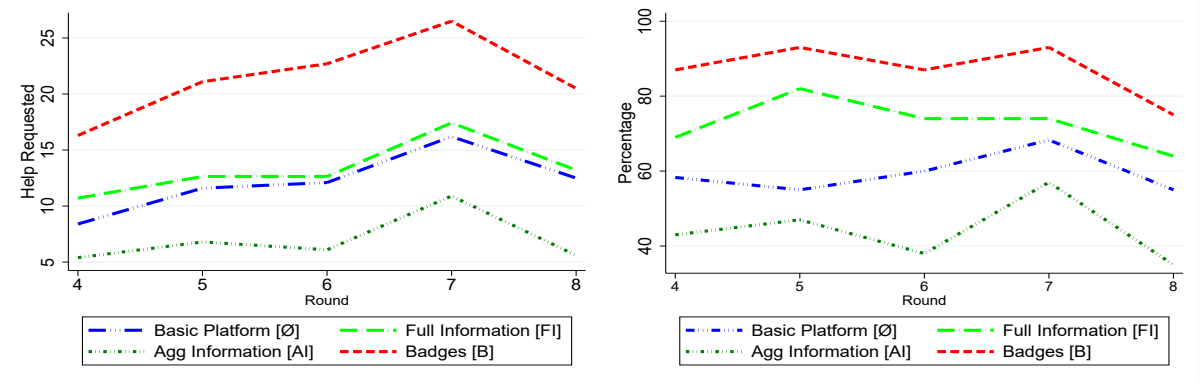
We are interested in seeing how the the amount of help requested (including help requests canceled) varies across the five rounds of the social media stage and if there are differences across treatments. In Figure A.1(a) and in Figure A.1(b), we plot the amount of help requested and the percentage of participants requesting help across round and across the four lab treatments. We notice that there is a general increase in the amount of help requested in all treatments in line with our results from Table A.6 discussed previously. Specifically, we see a similar percentage increase in the amount of help requested between Round 4 and Round 7 in the [Ø] and [B] treatments (78% versus 69%). Furthermore, as seen in Figure A.1(b), the percentage of participants requesting help remains fairly consistent across rounds (nptrend test, minimum p>0.2 even when last round is ignored). This shows that the effectiveness of the badges treatment is mainly due to the jump in the amount of helping behavior in the first round of the social media stage.

Table A.7: Help Given Based on Previous Help Received

VARIABLES	Basic Treatment	AI Treatment	FI Treatment	B Treatment
Individual Ability	0.117***	0.0655**	0.0694**	0.00718
	(0.0323)	(0.0317)	(0.0314)	(0.0596)
Previous Help Received	0.0816	0.0247	0.338***	0.0159
	(0.0911)	(0.0748)	(0.0494)	(0.0637)
Constant	-0.876**	-0.386	-0.292	3.139**
	(0.429)	(0.498)	(0.663)	(1.292)
Observations	240	240	336	240
Number of Subjects	60	60	84	60

Robust standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

Figure A.1: Helping Trends Across Rounds



(a) Average Help Requested Per Group

(b) % Participants Requesting Help

A.6 Further Analysis of Descriptive Social Norms in the [AI] Treatment

In the presence of descriptive social norms, we would expect an unhelpful participant in a helpful group to help more. The converse should also hold for helpful participant in an unhelpful group. We use non-parametric tests to compare helping behavior in the [AI]

treatment, where descriptive social norms are more likely to be present, to the [Ø] treatment, where descriptive social norms are unlikely to be present. Groups are categorized as helpful or unhelpful based on how their help given compares to the treatment average and the most and least helpful individuals are selected based on the average amount of help given across the five rounds of the social media stage. Our results show no difference in helping behavior in either of the two cases (rank sum test with, all $p > 0.2$). One limitation is that participants in our experiment remain with their group for all five rounds of the social media platform. As such, they do not experience the possibly different helping norms of different groups, which may otherwise impact their helping behavior. We conclude that, in our experiment, the overall helpfulness of the group is unlikely to alter the helping behavior of participants, which also contributes to the lack of effectiveness in the [AI] treatment.

A.7 Further Analysis of Reciprocity in the [FI] Treatment

Although our analysis shows that participants in the [FI] treatment are not targeting helpful individuals, they could be targeting individuals helpful to them. Specifically, a participant in the [FI] treatment can see the IDs of those who helped her, allowing her to reciprocate the help. If this was the case, we expect to see her answer more of her helpers' subsequent requests than she would in the [Ø] treatment. We test this by looking at the following measure: given that Participant A helped Participant B, what percent of eligible requests from A does B answer? Our results show, however, no significant difference between the [FI] and [Ø] treatments (62% versus 58%, with rank sum test, p=0.1). More generally, participants could be targeting specific subsets of the other group members. In this case, there would be a larger divergence between the amount of help given to the most helped versus least helped person in [FI] compared to [Ø]. However, even by this most general metric for targeting we see no significant difference between the two treatments (3.68 versus 3.33, with rank sum p>0.2). This gives us further evidence that participants in the [FI] treatment do not use helping information to target the amount of help they give, contributing to the lack of effectiveness in the [AI] treatment.

A.8 Panel Regressions on Performance in Study 2

We run a panel regression of participant performance on Study 2 treatment dummies, accounting for participant random effects and using standard errors clustered at the group

level. Our regression results are shown in Table A.8. We find that, when we account for individual ability in Column 2, all our goal and badges treatments significantly improve performance over the basic platform treatment. Furthermore, we find reduced significance in all treatment coefficients when we account for help requested in Column 3 and help received in Column 4. This indicates that, as in Study 1, the increase in performance across our Study 2 treatments is primarily driven by an increase in helping behavior over the basic platform treatment.

Table A.8: Panel Regression: Number of Questions Answered per Individual Across Treatments and Rounds

Variables	No Controls	w/ Indiv. Ability	w/ Indiv. Ability & Help Requested	w/ Indiv. Ability & Help Received
$\overline{G3}$ Treatment	0.205	1.106**	0.786*	0.843*
	(0.757)	(0.493)	(0.457)	(0.461)
$\overline{B3}$ Treatment	1.245*	1.438***	0.947**	0.983***
	(0.641)	(0.405)	(0.374)	(0.372)
$\overline{BE}$ Treatment	1.015*	1.069***	0.689*	0.715*
	(0.606)	(0.406)	(0.385)	(0.398)
$\overline{BE5}$ Treatment	1.297	1.164***	0.761*	0.808**
	(0.863)	(0.413)	(0.410)	(0.410)
$\overline{BE35}$ Treatment	0.448	1.461***	0.981*	0.994*
	(0.919)	(0.563)	(0.585)	(0.588)
Individual Ability		0.873***	0.866***	0.869***
		(0.0238)	(0.0239)	(0.0239)
Individual Help Requested			0.392***	
			(0.0497)	
Individual Help Received				0.403***
				(0.0496)
Round Controls	Yes	Yes	Yes	Yes
Constant	17.88***	2.286***	2.157***	2.131***
	(0.468)	(0.494)	(0.487)	(0.491)
Observations	2,070	2,070	2,070	2,070
Number of Subjects	414	414	414	414

Robust standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

Notes: The coefficients from running a panel regression with random effects and clustering at the group level are reported. The dependent variable is the performance of an individual in a round measured as the number of questions she answered correctly. Individual Ability measured as average round performance in individual stage.

APPENDIX B

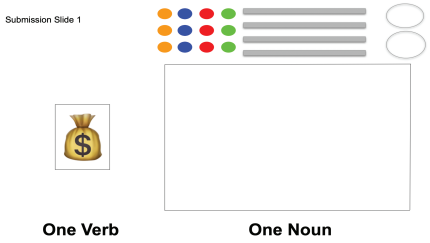
Appendix for "Creative Task Constraints and Knowledge Worker Productivity"

B.1 Emoji Set and Building Materials

Our experiment is implemented in Google Slides. Participants are given eight slides with each emoji already placed to the right and the drawing materials placed on top. Figure B.1a and Figure B.1b respectively display the emoji set and a practice slide.

Figure B.1: List of Emoji and Building Materials

(a) Emoji List

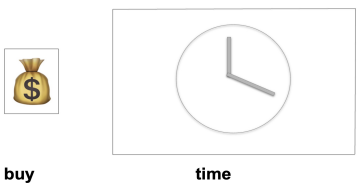


(b) Submission Slide

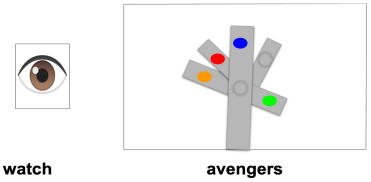
B.2 Image Examples

Participants in our experiment submit images with a wide range of originality and recognizability. In Figure B.2, we show an example of images with high originality and recognizability, high originality and low recognizability, low originality and high recognizability, and low originality and low recognizability. In (a), we see an example of a participant using the

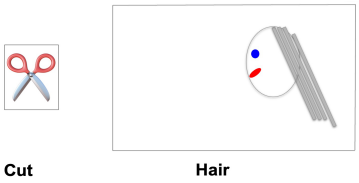
Figure B.2: Examples of Images with Varying Originality and Recognizability



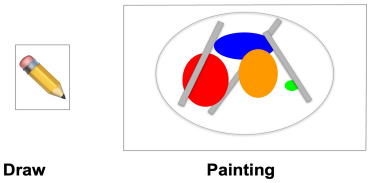
(a) High Originality and High Recognizablity



(b) High Originality and Low Recognizablity



(c) Low Originality and High Recognizablity



(d) Low Originality and Low Recognizablity

money emoji to represent an abstract concept such as "time," which is relatively uncommon in our experiment. The image comes from the T0 treatment, which has a 0% recognizability constraint, but the details the participant uses to make their image "stand out" also make it relatively recognizable to raters. Using the verb "buy" rather than "purchase" or "sell" helps the rater know that the noun might refer to an abstract notion of time associated with a clock rather than the clock itself (an alternative verb and noun could be "purchase clock"). In (b), the participant draws the hand of Thanos from the avengers movies. The image is unique in our experiment, but requires the raters having knowledge of the movie and understanding that this is a hand that possesses the "infinity stones" from the movie. Since this image is from a participant in the T80 treatment, we can assume that the participants was prioritizing originality and effectively giving up on passing the 80% constraint for this image. In (c), the participant draws an everyday object that can be quickly associated with the scissors emoji and includes enough details for the raters to know that the object is "hair." The image comes the T10 treatment and demonstrates an example of a simple recognizable image participants can create. Finally, in (d), the participant is unsuccessful in drawing a relatively common idea in our experiment such as "draw painting." The image comes from the T40 treatment and is an example of how participants can submit images that are creatively "inferior" to other images in our experiment.

B.3 Measuring Recognizability and Originality

Our experiment builds on the Consensual Assessment Technique (CAT) proposed by Amabile et al. (2018) to measure originality and the work proposed by Laske and Schroeder (2017) to measure recognizability.

B.3.1 Originality

The originality of an image is determined by our trained research assistants (RAs) that see the image's emoji, object, verb, and noun. Our research assistants are trained on a large set of practice images generated by participants in our pilot studies. They are given a broad description of the task, emoji, and building materials participants have access to, but do not know the purpose of our study, the recognizability of the images they rate, or which treatment they are coming from. Each research assistant works on sets of Qualtrics surveys that each have 32 images (the surveys are the same as the ones given to the Prolific raters to measure recognizability, but display each image's noun as well). RAs rate each image on 10 criteria according to a rubric we created based on the CAT, our insight, and discussions we had with the RAs. RAs give an image a rating from 1 to 5 on each criteria. For example, an image that receives a 5 out of 5 is one that the RA finds very original whereas one that receives a 1 out of 5 is not very original. Table B.1 provides a description of the 10 criteria and two flags that RAs rate each image on. Each image is rated by at least two research assistants and the average of the two ratings is used as the image's rating for that criteria.

B.3.2 Recognizability

The recognizability of an image is measured as the percentage of raters on Prolific (an online crowd-sourcing platform) that are able to guess the image's exact noun, seeing only its emoji, object, and verb. Prolific raters work through a Qualtrics survey that has 32 images consisting of the eight submitted images from four participants, each from one of the four treatments, placed in random order. Raters are given \$2.5 for completing the survey and receive \$0.03 for each noun they guess correctly. Before working on the survey, raters are informed that the images were created by participants in a decision making experiment, each image consists of an object a participant creates and an emoji representing an action on the object, and are shown the set of emoji and building materials the participants were given. Raters are informed that they need to write down exactly one word for the noun for each image. To make sure raters understand the instructions correctly, raters must answer three simple images correctly within three tries to continue to the survey. The images are simple

images for "car," "house," and "flower." Finally, we only allow each rater to complete one survey and raters only know the total number of nouns they answered correctly at the end of the survey. This makes sure that raters have a similar level of training when they guess the noun of all the images in our experiment.

B.4 Image Representationalism

While recognizability measures how accurately inexperienced raters can guess the object of an image (seeing its emoji, verb, and noun), representationalsim, determined by our judges, captures the strength of the relationship between an image's object and noun. This allows us to see if participants submit images with a clear connection or a "disconnect" between their idea and implementation, which can, at least indirectly, factor into their recognizability (for example, an image with a disconnect between idea and implementation is likely to be unrecognizable). We find a similarly high image representationalism across treatments (4.25, 4.34, 4.25, and 4.41 out of 5 in T0, T10, T40, and T80, ranksum test, minimum p=0.15). This shows that participants in all treatments submit images where the connection between the image's idea and its implementation is similarly clear to judges and suggests that, as expected, submitted images generally have, at least an indirect form of, recognizability factored in.

B.5 Treatment Ratings Across All Dimensions

In Table B.2 we show the results of our treatments across all the image rating dimensions. As expected, we find that the T10, T40, and T80 treatments generally result in ideas with lower novelty of idea and and novel use of materials resulting in a lower overall image originality. Interestingly, representationalism, expression, technical goodness, and complexity of implementation remain similar across all treatments, even in the T0 treatment where participants are paid regardless of an image's recognizability.

Table B.1: Description of Each Rating Criteria

Criteria	**Description**
Effort	Captures the amount of "effort" the participant placed in creating the image compared to the general amount of effort seen in other images submitted in the experiment
Representationalism	Captures the strength of the relationship between the object and noun (ignoring the emoji and verb)
Novel Use of Materials	Captures how frequently the image's building materials have been used separately and/or together to portray certain objects or ideas
Expression	Captures the strength of the relationship between the emoji and object and the verb and noun
Novelty of Idea	Captures how common the idea of the image is in our experiment (ignoring its implementation)
Liking	Captures how much the research assistant "likes" the image
Technical Goodness	Captures how well the idea of an image is implemented compared to other images with the same or similar ideas or themes
Complexity of Implementation	Captures the level of detail included in the image
Abstractness of Idea	Captures the abstractness of the image's idea
Originality	Captures the overall originality of the image, how "different" it is compared to other images in the experiment
Familiarity	A research assistant can use this criteria to "flag" an image they believe they do not have enough knowledge to rate (was never raised by research assistants)
Meeting Requirements	A research assistant can use this criteria to "flag" an image they believe does not meet the requirements of our experiment (for example, if the verb is completely unrelated to the emoji)

Table B.2: Summary of Criteria Ratings Across Treatments

Dimension	**T0**	**T10**	**T40**	**T80**
Effort	3.28	3.19	3.01	3.24
Representationalism	4.26	4.34	4.25	4.41
Novel Use of Materials	2.99	2.73	2.47***	2.77
Expression	4.32	4.35	4.21	4.34
Novelty of Idea	3.10	2.74**	2.56***	2.86
Liking	3.00	3.03	2.66**	2.91
Technical Goodness	3.77	3.71	3.51*	3.79
Complexity of Implementation	3.10	2.99	2.81	3.06
Abstractness of Idea	1.43	1.29*	1.34	1.40
Originality	3.06	2.73**	2.52***	2.81

BIBLIOGRAPHY

Acar, O. A., Tarakci, M., and Van Knippenberg, D. (2019). Creativity and innovation under constraints: A cross-disciplinary integrative review. *Journal of Management*, 45(1):96–121.

Allport, A., Styles, E., and Hsieh, S. (1994). Shifting intentional set: Exploring the dynamic of control tasks. *The MIT Press, Cambridge, MA*, pages 421–452.

Amabile, T. M., Collins, M. A., Conti, R., Phillips, E., Picariello, M., Ruscio, J., and Whitney, D. (2018). *Creativity in context: Update to the social psychology of creativity*. Routledge.

Amabile, T. M. and Pratt, M. G. (2016). The dynamic componential model of creativity and innovation in organizations: Making progress, making meaning. *Research in organizational behavior*, 36:157–183.

Anderson, A., Huttenlocher, D., Kleinberg, J., and Leskovec, J. (2013). Steering user behavior with badges. In *Proceedings of the 22nd International Conference on World Wide Web*, pages 95–106.

Baker, W. E. and Bulkley, N. (2014). Paying it forward vs. rewarding reputation: Mechanisms of generalized reciprocity. *Organization Science*, 25(5):1493–1510.

Bardhan, I. R., Krishnan, V. V., and Lin, S. (2007). Project performance and the enabling role of information technology: An exploratory study on the role of alignment. *Manufacturing & Service Operations Management*, 9(4):579–595.

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

Berg, J. M. (2014). The primal mark: How the beginning shapes the end in the development of creative ideas. *Organizational Behavior and Human Decision Processes*, 125(1):1–17.

Bista, S. K., Nepal, S., Colineau, N., and Paris, C. (2012). Using gamification in an online community. In *8th International Conference on Collaborative Computing: Networking, Applications and Worksharing*, pages 611–618. IEEE.

Bornstein, G. and Weisel, O. (2010). Punishment, cooperation, and cheater detection in "noisy" social exchange. *Games*, 1(1):18–33.

Bowden, E. M. and Jung-Beeman, M. (2003). Normative data for 144 compound remote associate problems. *Behavior research methods, instruments, & computers*, 35:634–639.

Bradler, C., Dur, R., Neckermann, S., and Non, A. (2016). Employee recognition and performance: A field experiment. *Management Science*, 62(11):3085–3099.

Brüggen, A., Feichter, C., and Williamson, M. G. (2018). The effect of input and output targets for routine tasks on creative task performance. *The Accounting Review*, 93(1):29–43.

Brzozowski, M. J., Sandholm, T., and Hogg, T. (2009). Effects of feedback and peer pressure on contributions to enterprise social media. In *Proceedings of the ACM 2009 International Conference on Supporting Group Work*, New York, NY, USA. Association for Computing Machinery.

Bughin, J. (2015). Taking the measure of the networked enterprise. *McKinsey Quarterly*, (4):19 – 22.

Bughin, J., Chui, M., Harrysson, M., and Lijek, S. (2017). Advanced social technologies and the future of collaboration. page *McKinsey Global Institute*.

Bulgurcu, B., Van Osch, W., and Kane, G. C. (2018). The rise of the promoters: user classes and contribution patterns in enterprise social media. *Journal of Management Information Systems*, 35(2):610–646.

Calzolari, G. and Nardotto, M. (2017). Effective reminders. *Management Science*, 63(9):2915–2932.

Cardon, P. W. and Marshall, B. (2015). The hype and reality of social media use for work collaboration and team communication. *International Journal of Business Communication*, 52(3):273–293.

Charki, M., Boukef, N., and Harrison, S. (2018). Maximizing the impact of enterprise social media. page *MIT Sloan Management Review*.

Chaudhuri, A. (2011). Sustaining cooperation in laboratory public goods experiments: a selective survey of the literature. *Experimental Economics*, 14(1):47–83.

Cialdini, R. B. (2007). Descriptive social norms as underappreciated sources of social control. *Psychometrika*, 72(2):263.

Colzato, L. S., Szapora, A., and Hommel, B. (2012). Meditate to create: the impact of focused-attention and open-monitoring training on convergent and divergent thinking. *Frontiers in psychology*, 3:116.

Crama, P., Sting, F. J., and Wu, Y. (2019). Encouraging help across projects. *Management Science*, 65(3):1408–1429.

de Vries, J., de Koster, R., and Stam, D. (2016). Aligning order picking methods, incentive systems, and regulatory focus to increase performance. *Production and Operations Management*, 25(8):1363–1376.

Deterding, S., Dixon, D., Khaled, R., and Nacke, L. (2011). From game design elements to gamefulness: defining "gamification". In *Proceedings of the 15th International Academic MindTrek Conference: Envisioning Future Media Environments*, pages 9–15.

Drucker, P. F. (1999). Knowledge-worker productivity: The biggest challenge. *California management review*, 41(2):79–94.

Duncker, K. and Lees, L. S. (1945). On problem-solving. *Psychological monographs*, 58(5):i.

Ebert, R. J. (1976). Aggregate planning with learning curve productivity. *Management Science*, 23(2):171–182.

Erez, M. and Kanfer, F. H. (1983). The role of goal acceptance in goal setting and task performance. *Academy of Management Review*, 8(3):454–463.

Erez, M. and Zidon, I. (1984). Effect of goal acceptance on the relationship of goal difficulty to performance. *Journal of Applied Psychology*, 69(1):69.

Fehr, E. and Gächter, S. (2000). Cooperation and punishment in public goods experiments. *The American Economic Review*, 90(4):980–994.

Fischbacher, U. (2007). Z-tree: Zurich toolbox for ready-made economic experiments. *Experimental Economics*, 10(2):171–178.

Fischbacher, U., Gächter, S., and Fehr, E. (2001). Are people conditionally cooperative? evidence from a public goods experiment. *Economics Letters*, 71(3):397–404.

Fujimoto, T. and Clark, K. B. (1991). Product development performance: Strategy, organization, and management in the world auto industry. pages *Harvard Business School Press, Boston, MA*.

Gallus, J. (2017). Fostering public good contributions with symbolic awards: A large-scale natural field experiment at wikipedia. *Management Science*, 63(12):3999–4015.

Gibbs, J. L., Rozaidi, N. A., and Eisenberg, J. (2013). Overcoming the "ideology of openness": Probing the affordances of social media for organizational knowledge sharing. *Journal of Computer-Mediated Communication*, 19(1):102–120.

Gilhooly, K. J. (2016). Incubation and intuition in creative problem solving. *Frontiers in psychology*, 7:1076.

Grant, S. and Betts, B. (2013). Encouraging user behaviour with achievements: an empirical study. In *2013 10th Working Conference on Mining Software Repositories (MSR)*, pages 65–68. IEEE.

Guilford, J. P. (1967). The nature of human intelligence.

Herroelen, W. and Leus, R. (2005). Project scheduling under uncertainty: Survey and research potentials. *European Journal of Operational Research*, 165(2):289 – 306.

Hopp, W. J., Iravani, S. M. R., and Liu, F. (2009). Managing white-collar work: An operations-oriented survey. *Production and Operations Management*, 18(1):1–32.

Hwang, E. H., Singh, P. V., and Argote, L. (2015). Knowledge sharing in online communities: Learning to cross geographic and hierarchical boundaries. *Organization Science*, 26(6):1593–1611.

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

Kagel, J. H. and Roth, A. E. (2016). *The Handbook of Experimental Economics*, volume 2 of *Economics Books*. Princeton University Press.

Keser, C. and Van Winden, F. (2000). Conditional cooperation and voluntary contributions to public goods. *scandinavian Journal of Economics*, 102(1):23–39.

Kiesel, A., Steinhauser, M., Wendt, M., Falkenstein, M., Jost, K., Philipp, A. M., and Koch, I. (2010). Control and interference in task switching—a review. *Psychological bulletin*, 136(5):849.

Kosfeld, M. and Neckermann, S. (2011). Getting more work for nothing? Symbolic awards and worker performance. *American Economic Journal: Microeconomics*, 3(3):86–99.

Laske, K. and Schroeder, M. (2017). Quantity, quality and originality: The effects of incentives on creativity.

Latham, G. P. and Locke, E. A. (2006). Enhancing the benefits and overcoming the pitfalls of goal setting. *Organizational Dynamics*, 35(4):332 – 340.

Leonardi, P. M., Huysman, M., and Steinfield, C. (2013). Enterprise social media: Definition, history, and prospects for the study of social technologies in organizations. *Journal of Computer-Mediated Communication*, 19(1):1–19.

Li, C. (2015). Why no one uses the corporate social network. *Harvard Business Review*, 87(1111):1–9.

Li, J., Leider, S., Beil, D., and Duenyas, I. (2021). Running online experiments using web-conferencing software. *Journal of the Economic Science Association*, 7(2):167–183.

Locke, E. A. and Latham, G. P. (2002). Building a practically useful theory of goal setting and task motivation: A 35-year odyssey. *American Psychologist*, 57(9):705.

Lu, J. G., Akinola, M., and Mason, M. F. (2017). "switching on" creativity: Task switching can increase creativity by reducing cognitive fixation. *Organizational Behavior and Human Decision Processes*, 139:63–75.

Madjar, N. and Shalley, C. E. (2008). Multiple tasks' and multiple goals' effect on creativity: Forced incubation or just a distraction? *Journal of Management*, 34(4):786–805.

Marder, A. (2015). Stack overflow badges and user behavior: an econometric approach. In *2015 IEEE/ACM 12th Working Conference on Mining Software Repositories*, pages 450–453. IEEE.

Mednick, S. (1962). The associative basis of the creative process. *Psychological review*, 69(3):220.

Miron-Spektor, E. and Beenen, G. (2015). Motivating creativity: The effects of sequential and simultaneous learning and performance achievement goals on product novelty and usefulness. *Organizational Behavior and Human Decision Processes*, 127:53–65.

Monsell, S. (2003). Task switching. *Trends in cognitive sciences*, 7(3):134–140.

Moreau, C. P. and Dahl, D. W. (2005). Designing the solution: The impact of constraints on consumers' creativity. *Journal of Consumer Research*, 32(1):13–22.

Mortensen, M. and Gardner, H. K. (2017). The overcommitted organization. *Harvard Business Review*, 95(5):58–65.

Napoleon, K. and Gaimon, C. (2004). The creation of output and quality in services: A framework to analyze information technology-worker systems. *Production and Operations Management*, 13(3):245–259.

Oktay, H., Taylor, B. J., and Jensen, D. D. (2010). Causal discovery in social media using quasi-experimental designs. In *Proceedings of the First Workshop on Social Media Analytics*, pages 1–9.

Ostermaier, A. and Uhl, M. (2020). Performance evaluation and creativity: Balancing originality and usefulness. *Journal of Behavioral and Experimental Economics*, 86:101552.

Ponsignon, F., Smart, P. A., and Maull, R. S. (2011). Service delivery system design: characteristics and contingencies. *International Journal of Operations & Production Management*, 31(3):324–349.

Rabin, M. (1993). Incorporating fairness into game theory and economics. *The American Economic Review*, 83(5):1281–1302.

Rode, H. (2016). To share or not to share: the effects of extrinsic and intrinsic motivations on knowledge-sharing in enterprise social media platforms. *Journal of Information Technology*, 31(2):152–165.

Roels, G. and Su, X. (2013). Optimal design of social comparison effects: Setting reference groups and reference points. *Management Science*, 60(3):606–627.

Runco, M. A. and Jaeger, G. J. (2012). The standard definition of creativity. *Creativity research journal*, 24(1):92–96.

Schlapp, J., Oraiopoulos, N., and Mak, V. (2015). Resource allocation decisions under imperfect evaluation and organizational dynamics. *Management Science*, 61(9):2139–2159.

Schmenner, R. W. (2015). The pursuit of productivity. *Production and Operations Management*, 24(2):341–350.

Schultz, K. L., Juran, D. C., and Boudreau, J. W. (1999). The effects of low inventory on the development of productivity norms. *Management Science*, 45(12):1664–1678.

Scopelliti, I., Cillo, P., Busacca, B., and Mazursky, D. (2014). How do financial constraints affect creativity? *Journal of Product Innovation Management*, 31(5):880–893.

Seaborn, K. and Fels, D. I. (2015). Gamification in theory and action: A survey. *International Journal of Human-computer Studies*, 74:14–31.

Shunko, M., Niederhoff, J., and Rosokha, Y. (2018). Humans are not machines: The behavioral impact of queueing design on service time. *Management Science*, 64(1):453–473.

Sio, U. N. and Ormerod, T. C. (2009). Does incubation enhance problem solving? a meta-analytic review. *Psychological bulletin*, 135(1):94.

Smith, R. D. and Robey, D. (1973). Research and applications in operations management: Discussion of a paradox. *The Academy of Management Journal*, 16(4):647–657.

Smith, S. M. and Blankenship, S. E. (1991). Incubation and the persistence of fixation in problem solving. *The American journal of psychology*, pages 61–87.

Smith, S. M., Gerkens, D. R., and Angello, G. (2017). Alternating incubation effects in the generation of category exemplars. *The Journal of Creative Behavior*, 51(2):95–106.

Song, H., Tucker, A. L., Murrell, K. L., and Vinson, D. R. (2018). Closing the productivity gap: Improving worker productivity through public relative performance feedback and validation of best practices. *Management Science*, 64(6):2628–2649.

Spohrer, J. and Maglio, P. P. (2008). The emergence of service science: Toward systematic service innovations to accelerate co-creation of value. *Production and Operations Management*, 17(3):238–246.

Srba, I. and Bielikova, M. (2016). A comprehensive survey and classification of approaches for community question answering. *ACM Transactions on the Web (TWEB)*, 10(3):1–63.

Stieglitz, S., Riemer, K., and Meske, C. (2014). Hierarchy or activity? the role of formal and informal influence in eliciting responses from enterprise social networks. In *ECIS*.

Stratman, J. K., Roth, A. V., and Gilland, W. G. (2004). The deployment of temporary production workers in assembly operations: a case study of the hidden costs of learning and forgetting. *Journal of Operations Management*, 21(6):689 – 707.

Strobach, T., Liepelt, R., Schubert, T., and Kiesel, A. (2012). Task switching: effects of practice on switch and mixing costs. *Psychological research*, 76:74–83.

Van Osch, W., Steinfield, C. W., and Balogh, B. A. (2015). Enterprise social media: Challenges and opportunities for organizational communication and collaboration. In *2015 48th Hawaii International Conference on System Sciences*, pages 763–772. IEEE.

Wu, C.-L., Huang, S.-Y., Chen, P.-Z., and Chen, H.-C. (2020). A systematic review of creativity-related studies applying the remote associates test from 2000 to 2019. *Frontiers in psychology*, 11:573432.