FACULTY OF ECONOMICS AND BUSINESS

THE CONSEQUENCES OF INCENTIVE PROGRAM DESIGN ON EMPLOYEE BEHAVIOR

Dissertation presented to
obtain the degree of
Doctor in Business Economics

by

Liliana Dewaele

Number 779 2022

Doctoral Committee

Advisors

Prof. Dr. Eddy Cardinaels	KU Leuven – Tilburg University
Prof. Dr. Alexandra Van den Abbeele	KU Leuven

Members

Prof. Dr. Markus Arnold	University of Bern
Prof. Dr. Sophie De Winne	KU Leuven
Prof. Dr. Christophe Feichter	WU Vienna
Prof. Dr. Sabra Khajehnejad	KU Leuven

Chairperson

Prof. Dr. Kristien Smedts	KU Leuven

Acknowledgments

Almost five years ago, I got an email from my supervisor prof. dr. Eddy Cardinaels, asking me if I was interested in pursuing a PhD. A few weeks after, I found myself working in Leuven's Hogenheuvelcollege. And now, almost five years later, I find myself writing this part of my dissertation. It has been quite a ride, and as such, I would like to thank the people who supported me along the way.

First and foremost, I would like to express my sincere gratitude to both my supervisors, prof. dr. Eddy Cardinaels and prof. dr. Alexandra Van den Abbeele. Eddy, thank you for always being available when I needed advice. Your support and encouragement during hard times motivated me to be persistent. Your detailed feedback, invaluable suggestions and comments contributed to a dissertation that I can be really proud of. Thank you for all the knowledge you shared with me. I can now agree with your belief that every piece of research is an incremental contribution to the literature and that eventually, we all contribute to novelty and knowledge creation. Alexandra, thank you for your guidance and words of wisdom throughout the entire process of writing the dissertation. Thank you for facilitating conference and course participations. You stimulated me to build a personal network in academia and I am truly grateful for that. Looking back, I realize I could not have been luckier to work with such passionate and skilled accounting researchers. Thank you both for always believing in me, and for giving me every opportunity to become the researcher that I am today.

Second, I want to thank Prof. dr. Sophie De Winne, Prof. dr. Sabra Kahjenehad, Prof. dr. Markus Arnold, Prof. dr. Christoph Feichter and Prof. dr. Kristien Smedts. Thank you for your willingness to take place on my dissertation committee. Sophie, thank you for your valuable feedback on all chapters of this dissertation. Your insights encouraged me to think about the bigger picture of my research. Sabra, thank you for your guidance and feedback on my work. Your methodological and analytical skills really improved the quality of my research projects. Markus and Christoph, I am deeply grateful for the time and effort you have

invested in my research. Receiving feedback from renowned researchers in performance evaluation was an honor. Kristien, thank you for chairing my committee, and for having given me the opportunity to execute pre-tests with your students back in 2019.

Furthermore, I would like to thank both Tilburg University and KU Leuven for funding data collection. This dissertation would not have been possible without this support.

Next, I would like to thank my former colleagues at AFI. You made hard times during the process of my PhD easier. So, thank you for listening to me when I was going through tough times. But also: thank you for taking my mind off the dissertation work every now and then. I look back at my time at KU Leuven enjoying our lunches, trips to the Ardennes, after-work drinks and even movie nights at the Hogenhovelcollege. I also want to thank all colleagues who have helped me in pre-testing my experiments, you know who you are. Finally, I want to thank my fellow final year PhDs. Liza, Carola, Maxim, Elien, Mathias, we started this journey together and we (almost) made it! Wishing you all the best of luck in the afterlife!

A few colleagues deserve special mention. Ditmir, Mathilde, Mieke and Jeffrey, thank you for helping me carrying boxes full of pre-tests through Leuven. Qinnan, we spent our first year together at the same office. Thank you for all the long work-related and not-so-work-related talks we had. It is truly amazing to have witnessed the making of such a great researcher! Mathilde, thank you for everything. You are my greatest supporter. Over the years of sharing a physical and even a virtual office, we became friends, keeping each other updated at all times. Marte, our time at the office together was short. Nevertheless, we succeeded in creating a strong bond with each other. I admire your persistence and the way you undertake research rigorously. I enjoyed our chats about oTree and I am always happy to see notifications in our python chat room. Elien, as a fellow experimentalist in management accounting we attended a lot of courses and conferences together. We always had great fun traveling together, being kicked out of our apartment in Amsterdam, and keeping each other awake during U.S. time zone conferences. I am looking forward to continuing to give each other peer feedback and to start writing that Voermans and Dewaele

paper! Finally, I want to thank Dieter. Even though we were not working together on research projects, I could always count on you if I needed help. Thank you for inspiring me and helping me with design choices whenever I had the feeling of being stuck.

October 2021 marked the date that I started a new job at the Open Universiteit, and as such, I would like to thank my new colleagues as well. Dennis and Willem, thank you for giving me the opportunity to work at the OU. Dennis, thank you for giving me the freedom to continue working on my PhD and for supporting me in my research endeavors. Frank, Thomas T, thanks both for helping me along the way, and for answering all my questions. Olga, thank you for becoming a friend, and for keeping me accountable for my PhD work! Apart from those already mentioned: Thomas P, Thomas VZ, Kenneth, Frederique, Bram, Alex, Stephanie, Bart, Sjuul, Han, Annelies, Karen, thank you for creating the nicest atmosphere in our department. It is a pleasure to work with you!

I would also like to thank my friends and family. Anneleen, Celine, Mahaut and Yndia, thank you for organizing all those awesome trips and dinners. These distractions were undoubtedly helpful in regaining energy to move forward with my dissertation. Mami, gracias por darme tu fuerza y por apoyarme en todo lo que hago. Papa, danku om een luisterend oor te zijn tijdens onze looptrainingen, en om te proberen begrijpen wat een PhD juist inhoudt. Thank you both for believing in me and for supporting me throughout this journey. Meter en Peter, bedankt om mij met open armen te ontvangen in Boezinge, waar ik steeds terecht kon om ongestoord te werken. Jullie zijn echt de allerbeste grootouders, jullie zorg en steun hebben ongetwijfeld geholpen in het afwerken van mijn doctoraat. Loís, thank you for reminding me of ‘mens sana in corpore sano’. Investing in a healthy lifestyle certainly contributed to my productivity levels. Merci Denise pour ton aide et ton soutien durant mon doctorat. Tes paroles de sagesse m’ont aidés à relativiser pendant les moments difficiles. Mark and Else, thank you for considering me as one of your own, and for helping me along the way.

Finally, I would like to thank my husband Boris. You were the one that encouraged me to pursue a PhD in the first place. However,

you were not always a big fan of all my complaints during the first couple of years of the PhD program. Nonetheless, I would like to thank you for celebrating every win with me, even the smaller ones. Thank you for helping me putting things in perspective, and for reminding me that I shouldn't take myself too seriously. Thank you for supporting me infinitely during hard times. Your advice, optimism, patience, understanding, and great cooking skills are all responsible for having finished this dissertation.

Liliana Dewaele
Leuven, June 2022

Table of Contents

List of Figures

List of Tables

General Introduction

In this PhD dissertation, I investigate the use and consequences of incentive programs on employee behavior from a management control perspective. The first section of this introduction describes the general research motivation. The second section motivates the application of the empirical research method to all three studies in this dissertation. Finally, the third section presents an overview of the chapters included in this dissertation.

Research motivation

Management control systems (MCS) are defined as tools used by organizations to guide employees to behave in line with the organizational goals (Merchant & Van der Stede, 2017). This dissertation focuses on incentive systems as an example of such MCS, and how their design can affect employee outcomes such as effort performance and creativity. Incentive systems are a form of results control, meaning that rewards are provided to employees based on certain performance or result achievements (Merchant & Van der Stede, 2017). That is, the design of incentive systems includes both the measurement and evaluation of employee performance and the provision of organizational rewards. As such, performance evaluation and rewards have been cited as necessary control tools for organizations by prior literature (Ferreira & Otley, 2009; Höppe & Moers, 2011).

Prior management accounting scholars have focused on studying how to design MCS to obtain the desired results (Malmi & Brown, 2008). More specifically, there is an important strand in the literature examining the effects of different incentive systems on outcomes such as employee performance (Hales, Wang, & Williamson, 2015; Hannan, Krishnan, & Newman, 2008; Newman & Tafkov, 2014; Tafkov, 2013), employee effort (Brown, Evans, Moser, & Presslee, 2022; Brüggen & Moers, 2007) but also on employee creativity (Chen, Williamson & Zou, 2012; Grabner, 2014; Kachelmeier, Reichert, & Williamson, 2008; Kachelmeier & Williamson, 2010). Despite the fact that research has devoted considerable attention to studying the effects of incentive system design on employee outcomes, and also the fact that well-known consulting firms are designing their own unique incentive systems

(Pfeiffer & Velthuis, 2009), we believe that this dissertation contributes to a better understanding of how incentive system design affects employee outcomes beyond those studied in prior research.

The literature identifies several design characteristics impacting incentive system effectiveness, ranging from including choices into incentive systems (Bareket-Bojmel, Hochman, & Ariely, 2017; Caza, McCarter, & Northcraft, 2015; Hales et al., 2015; Williams & Luthans, 1992), making information from performance evaluation systems transparent (Bol, Kramer, & Maas, 2016; Hannan, McPhee, Newman, & Tafkov, 2013; Maas & Van Rinsum, 2013; Tafkov, 2013), to including different types of evaluation formats into performance evaluation systems (Bently, 2019; Berger, Harbring, & Sliwka, 2013; Brutus & Donia, 2010; Cardinaels & Feichter, 2021; David, 2013 Lampe, Shäffer, & Schaupp, 2021).

Rewards

The first goal of this dissertation is to study the effect of having employees choose a reward as part of incentive system design, on their performance. Prior research in accounting has devoted considerable attention to studying the effect of different types of rewards on employee motivation and performance (Choi & Presslee, 2022; Heninger, Smith, & Wood, 2019; Kelly, Presslee, & Webb, 2017; Mitchell, Presslee, Schulz, & Webb, 2021; Presslee, Vance, & Webb, 2013). However, literature examining the degree of choice within given incentive contracts remains scarce, even though organizations are increasingly making use of so-called reward choices (Baeten & Verwaeren, 2012; Hillebrink, Schippers, van Doorne-Huiskes, & Peters, 2008; Vidal-Salazar, Cordón-Pozo, & José, 2016).

The general conclusion drawn from prior literature examining the performance effects different reward types is that tangible rewards account for higher performance improvements than cash rewards (Choi & Presslee, 2022; Heninger et al., 2019; Jeffrey, 2009; Kelly et al., 2017).[1] Nevertheless, both types of rewards have been showed to be stored into different mental accounts (Choi & Presslee, 2022; Thaler, 1999). Hence, a better understanding of how reward choice affects

[1] Tangible rewards are defined as noncash rewards that have monetary value and are restricted in use (Choi & Presslee, 2022; Presslee et al., 2013).

employee performance is warranted. To this end, it is important to study the effect of reward choice by keeping the mental account of all included options constant, which brings out the motivation of the first chapter of this dissertation.

Peer performance evaluations

The second goal of this dissertation is to broaden our knowledge of the use of peer evaluations and their effects in the workplace. Substantial literature documents evidence of the effects of performance evaluations on employee outcomes thereby taking into account several evaluation design characteristics (Berger et al., 2013; Bol et al., 2016; Cardinaels & Feichter, 2021; Moers, 2005). However, the studies to date primarily focus on performance evaluations done by managers, and their effects on employee performance. Nevertheless, prior literature documents that managerial performance evaluations are often too lenient (Bol et al., 2016; Kampkötter & Sliwka, 2011; Moers, 2005; Rynes, Gerhart, & Parks, 2005), which in turn causes harmful effects on employee motivation and performance (Bol, 2011; Moers, 2005; Prendergast, 1999).

As such, scholars have proposed to include multiple sources (managers, supervisors, peers, self, subordinates) in performance evaluations as a way to draw a more accurate picture of an employee's performance (Conway, & Huffcutt, 1997; Dalla Via, Hartmann, & Collini; 2012). While accounting scholars have examined the effects of managerial evaluations on employee outcomes, little is known about the effects of peer evaluations in particular. Some studies have examined different types of peer evaluations, and conclude that it can be a powerful control tool (Arnold, Hannan, & Tafkov, 2018, 2019; Towry, 2003), however evidence on the effect of the design of such peer evaluations is lacking (Jackson, Michaelides, Dewberry, Schwencke, & Toms, 2020).

This dissertation aims to provide further insights into how the format of peer evaluations can affect employee outcomes such as effort, acceptability, and creativity. That is, organizations are increasingly making use of peer evaluation systems as a monitoring tool (Holderness, Olsen, & Thornock ; 2017), even though research on how to design and implement these controls is limited. Indeed, organizations differ in how they design peer evaluations. Some organizations use peer ratings

(Homem de Mello, 2019), while others include narrative comments (Gorman, Meriac, Roch, Ray, & Gamble; 2017). And where some firms decide to display performance evaluation information to all employees, others decide to keep this information private (Hannan et al., 2013; Tafkov, 2013). All these different design formats of peer evaluations in turn, can impact employee behavior not only in the positive sense, but also in the negative sense (Carpenter, Matthews, & Schirm, 2010). Hence, the second and third chapter of this dissertation aim to study and understand how peer evaluation design affects employee behavior.

Method: laboratory experiments

In all three chapters of this dissertation, laboratory experiments are employed as the method to examine the research questions. Experiments involve interventions in data collection, where researchers elicit dependent variables, control the research setting, and purposefully vary or manipulate the independent variables (Bloomfield, Nelson, & Soltes, 2016). Experiments have the advantage to study the effects of management control interventions in a controlled environment (Falk & Heckman, 2009). That is, experiments provide a clean test of theory, by isolating the variables of interest and by excluding confounding variables. In addition to providing controlled variation, experiments also allow for inferring causal relationships, because they are characterized by high internal validity (Bloomfield et al., 2016; Kadous & Zhou, 2016). Finally, experiments allow for measuring dependent variables such as employee effort and performance more precisely as compared to observing them in the field (Sprinkle, 2003).

The studies presented in this dissertation benefit from the experimental method because it allows for studying whether and how managerial accounting practices affect the behavior of individuals (Sprinkle, 2003). In the first chapter, for example, we study how a mere reward choice affects employees' cognitive performance. Without the use of a laboratory experiment, it would have been difficult to observe and analyze the effects of a reward choice containing options from the same mental account, as organizations usually include a myriad of reward types into their flexible benefit plans. In the second and third chapters then, we study the effects of peer evaluation system design on

a number of outcomes that are difficult to measure precisely in organizational contexts. In organizational contexts, variables such as effort and creativity are often also affected by other variables than peer evaluation system design. Hence, the use of experiments allows us to document the causal effect of several peer evaluation designs on employee behavior, which can benefit organizations that are considering implementing or updating such control tools.

Participants in the experiments presented in this dissertation were either business students or Amazon mTurk workers. As the studies in this dissertation test behavioral theories, participants are not required to possess expert knowledge or considerable practical experience. Hence, the use of non-sophisticated participants to examine our research goals and objectives is warranted (Bloomfield et al., 2016).

Overview of the chapters

This dissertation consists of three chapters in which three experimental studies are presented that examine the effects of incentive system design on employee behavior. Each of the three chapters has been written as an independent scientific article. Therefore, it is possible that the research motivation and literature discussed in the separate chapters might exhibit some overlap. A visual representation of the three chapters can be found in Figure 1.

Chapter 1

The first chapter is co-authored with Eddy Cardinaels and Alexandra Van den Abbeele, and studies the effect of an employee reward choice on cognitive task performance. Prior research has primarily examined the effect of different types of rewards on employee motivation and performance (Choi & Presslee, 2022; Kelly et al., 2017; Mitchell et al., 2021; Presslee et al., 2013). However, these studies do not take into account the fact that organizations nowadays offer their employees a choice with respect to rewards (Heninger et al., 2019). Moreover, prior research shows that different types of rewards are typically stored in different mental accounts (Choi & Presslee, 2022; Thaler, 1999). As such, this study attempts to uncover the effects of a mere reward choice, by including only tangible rewards as options, as to keep the mental

account constant. Additionally, prior research shows that performance on more demanding task types is hard to motivate by financial incentives (Bonner, Hastie, Sprinkle, & Young, 2000). Hence, in this study we examine whether a reward choice can positively affect cognitive task performance.

We predict that having a reward choice can positively affect employees' cognitive task performance. We base our predictions on traditional economic cost-benefit analysis (Kool & Botvinick, 2018) by arguing that individuals will derive utility from choosing a tangible reward, as compared to when they are simply assigned one (Kube, Marchal, & Puppe, 2012). We thus expect that offering a reward choice increases an employee's subjective value of the reward, thereby increasing the benefits of exerting costly cognitive effort. Consistent with our predictions, our results show that individuals are incentivized to exert higher cognitive effort when they are offered a reward choice, compared to when no reward choice is offered. Our findings also show that a reward choice can only incentivize cognitive performance when its option set is large enough.

With this paper we contribute to the literature in several ways. First, we contribute to the scarce research base that documents positive effects of a reward choice on employee performance (Bareket-Bojmel et al., 2017; Caza et al., 2015; Williams & Luthans, 1992) by showing that these effects can also be extended to cognitive performance (Bonner et al., 2000). Second, we provide important insights into the effect of a reward choice containing options from the same mental account. As such, our findings make an important contribution to prior accounting literature examining the effects of different reward types, that have been shown to be stored into different mental accounts (Bareket-Bojmel et al., 2017; Choi & Presslee, 2022; Dzuranin et al., 2013; Kelly et al., 2017; Mitchell et al., 2021). Finally, our results contribute to a better understanding of how reward choices can promote cognitive performance in practice. More specifically, organizations employing individuals that have a cognitively demanding job, may benefit from being rewarded with a reward choice.

Chapter 2

The second chapter is co-authored with Eddy Cardinaels and Alexandra

Van den Abbeele, and studies the effect of transparency for given peer evaluation systems on employee effort. Traditional agency theory postulates that peer evaluations have beneficial effects on employee effort as they reduce free-riding in teams through increased communication (Erez, Lepine, & Elms, 2002; Marx & Squintani, 2009; Sol, 2016; Towry, 2003). However, this is in sharp contrast with what behavioral scientists find (Balafoutas, Czermak, Eulerich, Fornwagner, 2020; Carpenter et al., 2010). In this paper, we aim to replicate and extend the findings from prior literature. We do this by taking into account outcome transparency, as prior researchers often argue that individuals behave differently when information becomes publicly available (Bol et al., 2016; Hannan et al., 2013; Maas & Van Rinsum, 2013; Tafkov, 2013).

We collect data through an experiment in which participants are compensated based on a tournament incentive scheme. Lazear (1989) argues that individuals are motivated to win such competitions and that they can do this by either increasing their own effort or by harming others. We build our predictions based on self-concept maintenance theory, which postulates that individuals face an ethical dilemma when attempting to promote one's relative position in a competition (Mazar, Amir, & Ariely, 2008). We predict and find that employee effort is affected by peer evaluation outcome transparency, for given peer evaluation systems in competitive settings.

Our findings are important for several reasons. We show that the use of an appropriate peer evaluation system design (i.e. peer rankings) can mitigate the disincentivizing effect of peer evaluations on employee effort found by prior literature when its outcomes are kept private (Carpenter et al., 2010). Additionally, we show that making peer evaluation outcomes transparent, peer evaluations consisting of ratings can increase employee effort more than when they consist of peer rankings. With these results, we extend prior management accounting literature by showing that the effects studied in traditional manager-employee settings translate to peer settings as well (Berger et al., 2013; Cardinaels & Feichter, 2021; Evans, Moser, Newman & Stikeleather, 2026; Maas & Van Rinsum, 2013).

Chapter 3

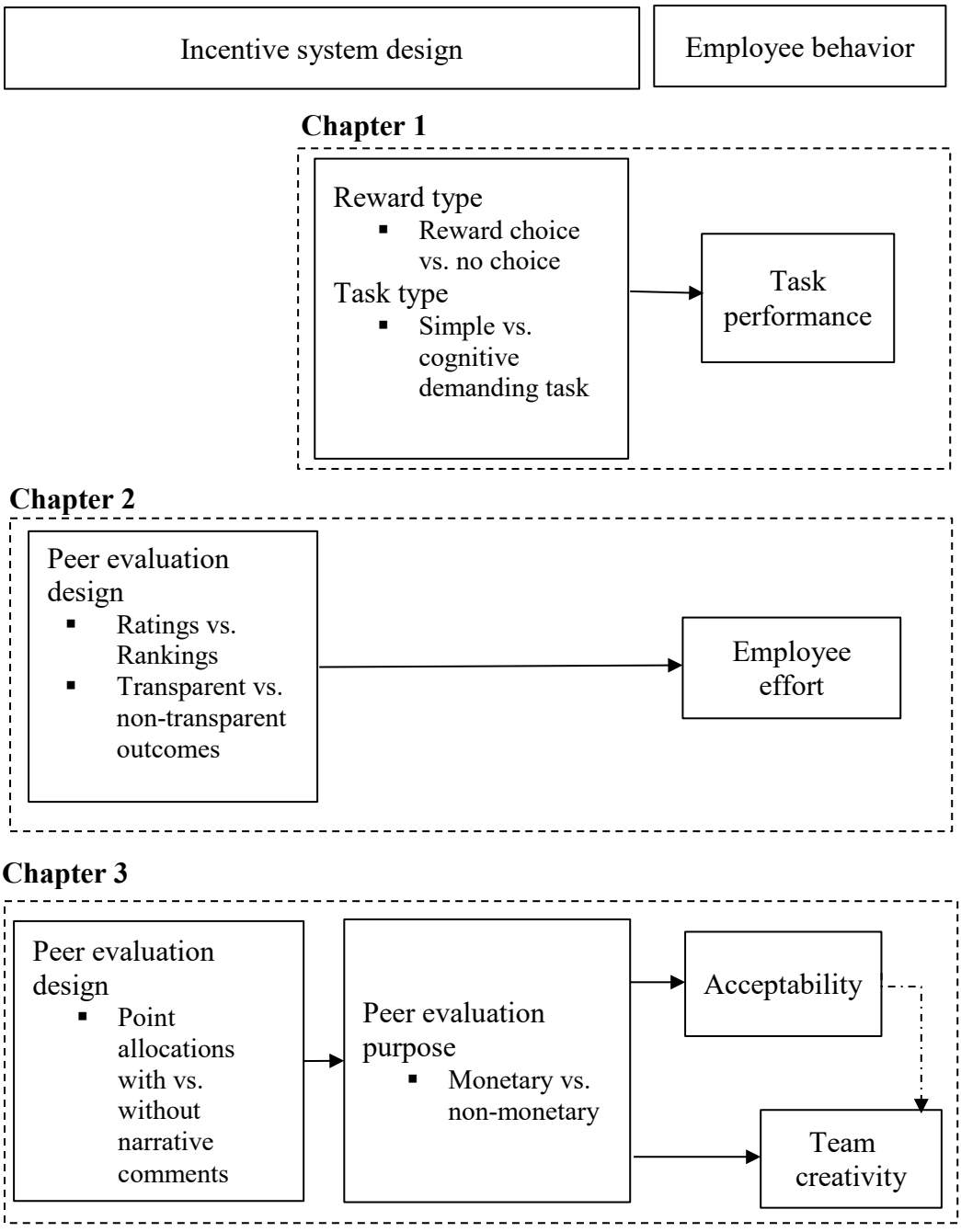
The third chapter is co-authored with Eddy Cardinaels and Alexandra Van den Abbeele, and studies how the role of narrative comments in peer evaluations influences acceptability and team creativity when managers use the evaluations for monetary or non-monetary purposes. After all, understanding how control systems affect employee acceptability of those systems, is an important research issue (Maley, Dabic, & Moeller, 2020). That is, the level of acceptability of such systems can be a potential driver for employee behavior (Ilgen, Fisher, & Taylor, 1979; Ilgen, & Davis, 2000). To this end, we study whether the inclusion of narrative comments can affect acceptability, in a team creative context, by taking into account different purposes for which these peer evaluations are used by managers. Indeed, researchers suggest that the purpose for which evaluations are being used, determines employee behavior in important ways (Appelo, 2015; Arnold et al., 2018, 2019; Brutus, 2010).

In order to examine the effect of peer evaluation design on acceptability and team creativity, we develop an experiment by manipulating the purpose for which peer evaluations are used (monetary and non-monetary) and whether or not peer evaluations include narrative comments. We predict and find that the inclusion of narrative comments increases acceptability levels, especially when managers use them for monetary purposes (i.e. bonus allocation). We further argue and observe that individuals engage in impression management (Ariely, Bracha, & Meier, 2009) when there is money at stake. Alternatively, we predict and find that the inclusion of narrative comments can only increase team performance when managers use peer evaluation outcomes in a more developmental way, by not tying explicit monetary bonuses to it.

We contribute to the existing literature on performance evaluations by showing that the use of narrative comments in peer evaluations is useful, but that they can also have unintended effects on creative performance. We thus contribute to the scarce literature on narrative feedback as a control tool by increasing our understanding of the role of narrative comments in performance evaluations (Arnold et al., 2018, 2019; Brutus, 2010; David, 2013; Lampe et al., 2021; Stubbs, 2021). Finally, with this paper, we make an important contribution to the literature on management controls in creative task settings, by focusing

on team creative performance rather than individual creative performance (Cardinaels, & Feichter, 2021; Cardinaels et al., 2021; Chen et al., 2012 Kachelmeier et al., 2008).

Figure 0.1: Overview of the chapters



Chapter 1

Too many choices: the effect of a reward choice on cognitive task performance*

Abstract:

This study examines the effect of reward choice in task settings that require different levels of cognitive resources from participants. Based on insights from behavioral economics we argue that reward choice and task type interact such that reward choice increases task performance more in tasks that are cognitively demanding compared to tasks that are simpler. To test this, we employ a laboratory experiment, manipulating reward type (i.e. reward choice versus no choice), and task type (i.e. simple versus more demanding task). Our results contribute to the literature on incentives by demonstrating that reward choice can increase task performance in cognitively demanding tasks. However, managers should carefully interpret these results as we also find a potential downside from offering reward choices in these more demanding tasks, which can result in slightly lower task persistence. Additionally, we find that the effect of reward choice on task performance is moderated by individual differences in personal growth initiative. More specifically, reward choice mainly leads individuals who score higher on personal growth initiative to increase performance in cognitively demanding tasks.

* This chapter is joint work with Eddy Cardinaels and Alexandra Van den Abbeele. We thank Markus Arnold, Willie Choi, Christoph Feichter, Sophie De Winne, Sabra Khajehnejad, Jason Kuang (discussant), seminar participants at KU Leuven, anonymous reviewers and conference participants at the 43rd EAA Annual Congress 2021, Management Accounting Section Midyear Meeting 2020, the Annual Conference for Management Accounting Research 2020, as well as faculty and participants at the EAA's Doctoral Colloquium in Accounting 2019 and at the ENROAC Doctoral Summer School in Management Accounting 2018.

1.1 Introduction

This paper examines the effect of a tangible reward choice and task type on cognitive task performance. While many papers have focused on different types of rewards and their effect on motivation and performance (Choi & Presslee, 2022; Heninger, Smith, & Wood, 2019; Kelly, Presslee, & Webb, 2017; Mitchell, Presslee, Schulz, & Webb, 2021; Presslee, Vance, & Webb, 2013), we study the effect of a reward choice among hedonic tangible rewards. Moreover, we study the effect of a tangible reward choice on cognitive task performance, as prior literature shows that performance on more demanding task types is hard to motivate by financial incentives (Bonner, Hastie, Sprinkle, & Young, 2000).

The focus on reward choice among hedonic tangible rewards is adopted for three reasons. First, the use of tangible rewards in the form of noncash incentives with monetary value which are restricted in use (Presslee et al., 2013), has become increasingly common in practice (Haden, 2017; Hammermann & Mohnen, 2014; Incentive Research Foundation, 2017; Peltier, Schultz, & Block, 2005; People Driven Performance, 2015). A survey by the Incentive Federation (2016) in the US reports that 36% and 71% of employees receive points systems and gift cards, respectively. This evidence shows that the use of reward choice, is quite prominent among incentive systems used by employers today (Baeten & Verwaeren, 2012; Hillebrink, Schippers, van Doorne-Huiskes, & Peters, 2008; Vidal-Salazar, Cordón-Pozo, & José, 2016). Furthermore, a large portion of tangible rewards used in practice are hedonic in nature (Incentive Research Foundation, 2018). Second, prior evidence shows that tangible rewards lead to superior performance compared to cash rewards (Choi & Presslee, 2022; Dzuranin, Randolph, & Stuart, 2013; Heninger et al., 2019; Jeffrey, 2009; Kelly et al., 2017; Shaffer & Arkes, 2009). Recent findings also indicate that hedonic tangible rewards have a larger motivating potential than utilitarian tangible rewards, because the former are more likely to be stored in a different mental account other than salary or cash (Choi & Presslee, 2022; Mitchell et al., 2021). Finally, evidence indicates that individuals prefer hedonic tangible rewards over cash rewards, which makes our reward choice attractive to individuals (Shaffer & Arkes, 2009).

While research both in the field of social psychology and consumer behavior, has shown that making choices can be difficult (Iyengar & Lepper, 2000; Patall, Cooper, & Robinson, 2008; Sela, Berger, & Liu, 2009), some scholars argue that the effect of reward choice can spark motivation (Bareket-Bojmel, Hochman, & Ariely, 2017; Caza, McCarter, & Northcraft, 2015). We use insights from behavioral economics, and more specifically the concept of subjective valuation, to predict an interaction effect between reward type (i.e. having a reward choice or not) and task type (i.e. a simple task versus a more demanding task). To predict the subjective value of a reward, we rely on insights from gift-exchange literature which postulates that individuals can derive utility from matching a chosen option to their preferences (Kube, Marchal, & Puppe, 2012). As such, we expect that offering a reward choice to individuals increases the subjective value of the reward (due to preference matching) in cognitively demanding tasks, which could eventually make exerting costly cognitive effort more attractive.

To test our predictions, we conduct a 2 x 2 laboratory experiment, where we manipulate (1) reward type (no choice versus reward choice) and (2) task type (simple task versus more demanding task). Participants first receive information on their payoffs, i.e. a non-task contingent tangible reward they receive for participating in the study, whereby one group received a random reward, while the other group could explicitly choose among an extended range of options. That is, individuals in this study operate in a gift-exchange setting. After the reward was assigned or chosen, participants worked on a task that was comparable in nature (highlighting the letter "e" in a given text box – adapted from Baumeister et al. 1998). However, depending on the assigned condition, the task differed in the extent to which cognitive effort was required to solve the task.

Our results show a significant interaction effect of reward type and task type. Consistent with our prediction, the results suggest that choosing from an extensive choice set of rewards followed by a cognitive effortful task, incentivizes higher performance compared to when participants do not receive the option to choose. These results suggest that reward choice can effectively incentivize higher performance in a more cognitive demanding task. As a validation of our theory we also show that positive effects of reward choice on task performance in demanding tasks is higher for individuals who score higher on personal

growth initiative. However, our results also show that reward choice can have a potential downside suggesting that individuals work less persistent in cognitively demanding tasks when combined with reward choice. Nevertheless, this quitting behavior does not reduce total effort expended on the task. Finally, we also show that the positive effect of reward choice in demanding tasks only materializes when the option set is large enough as our additional results highlight that offering a limited reward choice including three options produces similar effects as having no choice.

Our study makes four main contributions to the literature. First, few studies have examined the effects of reward choices given to individuals (Bareket-Bojmel et al., 2017; Caza et al., 2015; Williams & Luthans, 1992). Our study fills this gap by explicitly examining the effects of reward choice in a laboratory experimental setting. Earlier management accounting studies have primarily focused on the different effects of cash rewards versus noncash (i.e. tangible) rewards (Bareket-Bojmel et al., 2017; Choi & Presslee, 2022; Dzuranin et al., 2013; Kelly et al., 2017; Mitchell et al., 2021). A notable exception is the study by Heninger et al. (2019) which examines university employees' reward type preferences, by administering a choice between gift cards, cash and tangible rewards. Nevertheless, the authors study performance effects of a university's wellness program under the different types of self-selected rewards rather than the effect of the reward choice itself. We add to this stream of literature by focusing on a reward choice consisting of tangible rewards only, thereby keeping the mental account of the rewards constant.[2] We show beneficial effects of reward choice in cognitively demanding tasks, but only when participants are offered an extensive set of choices.

Second, we add to earlier research in examining the effect of employee participation in incentive program design (Cox, 2000; Groen, 2018; Lawler & Hackman, 1969; Venkatesh & Blaskovich, 2012; Wong-On-Wing, Guo & Lui, 2010). This literature stream finds that having employees participate in the incentive program design, increases their fairness perception which is in turn beneficial for employee performance. With this study, we add to this stream of literature by showing that

[2] Mental accounting theory posits that rewards are stored into different mental accounts, such that salary is stored in a mental account associated to more utilitarian expenses, whereas tangible rewards are stored in a different mental account associated to more hedonic expenses (Choi & Presslee, 2022; Thaler, 1999).

employee participation in incentive programs has the potential to improve employee performance in demanding tasks by offering them reward choices.

Third, we add to the management accounting literature by suggesting that certain types of incentives, in particular reward choices, are effective in improving performance in cognitive demanding tasks (Bonner et al., 2000). Our results, however, need to be taken with some caution. While we find that reward choice is beneficial to task performance in complex tasks, it might still affect persistency in that individuals might stop working on the task slightly earlier than counterparts that are not offered a reward choice.

Finally, our findings are important from a practical perspective. The use of reward choices becomes more popular in organizations based on the believe that a diverse set of rewards may help to spark motivation among employees. Next to institutional pressures such as the legal and fiscal framework concerning employee compensation, efficiency gains (i.e. gains in employee performance) are often cited as another major advantage of offering reward choice (Baeten & Verwaeren, 2012). With this paper, we focus on the efficiency gains that could be reaped by using reward choices as a flexible benefit system. More specifically, we provide knowledge on the pros and cons of using a reward choice to compensate individuals. We broaden the understanding of the effects of reward choice by investigating under which circumstances they work. Our results shed light on the interactive effects of reward choice and cognitive costly tasks, indicating that the effect of reward choice depends on the nature of the job. More specifically, reward choice compared to no choice increases a participant's performance only in more demanding tasks. We also show that, the option set should be large enough for reward choice to have a positive effect on performance in this more demanding task. Finally, our results indicate that the effect of reward choice is minimal in simple tasks.

1.2 Background and hypothesis development

1.2.1 Background

1.2.1.1 Task type

Accounting researchers have traditionally sought to understand how task performance can be improved through the use of incentives (Bonner et al., 2000; Bonner & Sprinkle, 2002; Choi, Clark, & Presslee, 2019). As such, Bonner and colleagues (2000) have established that financial incentives are less likely to improve performance as tasks become more cognitively demanding. More interestingly, Choi and colleagues (2019) find that incentive effects even differ across various task types that are not overly complex. As such, the authors find that fixed pay outperforms piece-rate pay in a letter search task (Kachelmeier, Thornock, & Williamson, 2016; Sprinkle, Williamson, & Upton, 2008), but not in a decoding task (Chow, 1983; Church, Libby, & Zhang, 2008; Fisher, Maines, Peffer, & Sprinkle, 2002) and a slider task (Chan, 2018; Gill & Prowse, 2013). Indeed, as argued by Bonner and colleagues (2000), performance across tasks that vary in type do not always elicit the same level of individual effort.

In the current study, we examine the interactive effect of reward choice and task type on performance. We follow prior accounting scholars in defining the more demanding version of the task in this study, as a task in which the component complexity differs from the easier task (Bonner et al., 2000; Wood, 1986). Wood (1986) argues that as the number of subtasks and cognitive acts increases, the component complexity of a task increases. That is, the more demanding version of the letter detection task by Baumeister and colleagues (1998) specifies individuals to take into account two task rules. In the simpler version of the task, individuals only take into account one rule, and their performance is a mere function of motivation and effort, rather than the use of cognitive strategies (Locke & Latham, 1990). Indeed, prior literature has documented lower task performance on more demanding versions of the letter detection task, compared to the easier version (Arber et al., 2017; Baumeister, Bratslavsky, Muraven, & Tice, 1998; Tice, Baumeister, Shmueli, & Muraven, 2007). As explained below, we predict

that a tangible reward choice incentive will have a motivating effect in more demanding tasks.

1.2.1.2 Tangible rewards

Tangible rewards are defined as noncash rewards that have monetary value and are restricted in use (Choi & Presslee, 2022; Presslee et al., 2013). These types of rewards are increasingly being offered to employees in today's business environment (Incentive Federation Inc., 2016; Incentive Research Foundation, 2015). More specifically, U.S. firms invest between $76.9 billion and $100 billion on tangible rewards, per year (Incentive Federation Inc., 2016; Incentive Research Foundation, 2018). The use of tangible rewards, and more specifically the choice herein, can take a variety of forms in practice. For example, firms commonly utilize points systems in incentive programs, in which employees can accumulate points that can later be redeemed for a hedonic tangible reward of choice (Alonzo, 1996; Jeffrey & Shaffer, 2007; Norberg, 2017).

The substantial use and economic importance of tangible rewards in practice have led scholars to examine its motivational effects on employee performance (and related outcomes). As such, a growing body of research finds that tangible rewards mainly affect performance positively (Choi & Presslee, 2022; Heninger et al., 2019; Jeffrey, 2009; Kelly et al., 2017; Mitchell et al., 2021; Presslee et al., 2013; Shaffer & Arkes, 2009). Prior studies in management accounting find that tangible rewards can increase performance compared to cash rewards (Choi & Presslee, 2022; Heninger et al., 2019; Jeffrey, 2009; Kelly et al., 2017).[3] The latter studies draw on mental accounting theory (Thaler, 1999) for predicting that tangible rewards are associated to a mental account distinct from salary, which affects effort positively compared to cash rewards. In a more recent study by Mitchell et al. (2021), the authors similarly draw on mental accounting theory to show that the nature of tangible rewards affects effort differently. The authors find that even tangible rewards are stored into different mental accounts according to

[3] Heninger and colleagues (2019) argue and find that gift cards are more incentivizing for performance in a university's wellness program than other tangible rewards. However, in this study we do not find differences across average performance for individuals rewarded with a gift card compared to individuals rewarded with other types of tangible rewards (such as experiences and merchandise).

their hedonic or utilitarian nature, and that these differences translate into performance differences. Similarly Choi and Presslee (2022) show that hedonic tangible rewards induce a larger amount of effort among individuals than utilitarian framed cash rewards. Finally, Presslee and colleagues (2013) find that call center employees rewarded with a cash reward outperformed those rewarded a tangible reward, because employees offered a tangible reward type set easier goals for themselves as compared to employees rewarded with cash (Presslee et al., 2013). While this body of research mainly investigates the effects of tangible rewards compared to cash rewards, research on how the choice between tangible rewards affects performance, remains largely unanswered.

Moreover, firms are increasingly offering employees flexible benefit plans (Baeten & Verwaeren, 2012; de la Torre-Ruiz, Vidal-Salazar, & Cordón-Pozo, 2019). These flexible benefit systems are defined as systems in which employees have a degree of freedom in selecting their benefits or rewards (Vidal-Salazar et al., 2016). Firms started offering these reward schemes to respond to their employees' personalized needs (Barringer & Milkovich, 1998; Perkins & Jones, 2020). Thus, given that both firms are increasingly rewarding their employees with tangible rewards and offering them choices over rewards, this study investigates the effect of a tangible reward choice.

1.2.1.3 Reward choice

There are three main strands in the literature on the construct of reward choice. A first body of research focuses on the effects of offering flexible benefit plans to employees, in which employees can choose their benefits or rewards (Barber, Dunham, & Formisano, 1992; Barringer & Milkovich, 1998). This line of research sheds light on the positive effects of flexible benefit plans on organizational outcomes such as increased company attractiveness and employee retention (Baeten & Verwaeren, 2012; Koo, 2011), culture alignment (Chiang & Birtch, 2006; Fay & Thompson, 2001), higher perceptions of procedural justice among employees (Cole & Flint, 2004) and higher perceptions of job quality among employees (Kelliher & Anderson, 2008).

A second body of research in turn, focuses on the performance effects of having employees design their own reward system. The findings suggest that worker performance only improves when workers make an actual reward choice rather than participating in the design of their

compensation package (Leana, Locke, & Schweiger, 1990; Morgeson, Campion, & Maertz, 2001).[4] Similar findings have emerged from prior management accounting research, where scholars shed light on a performance improvement effect of compensation package choice due to individuals self-selecting in the contract type that best matches their skills and/or personality) (Cardinaels, Chen, & Yin, 2018; Chow, 1983; Hales, Wang, & Williamson, 2015).

A third body of research focuses on the effects of reward choices on employee performance but does not hold the type of reward constant in the construct of reward choice. In a study by Williams and Luthans (1992), for instance, participants in the experiment can choose for an activity reward choice (i.e. rest periods during task execution) or an outcome reward choice (time off versus bonus pay). The authors show that reward choice increases performance, and that reward choice interacts with feedback such that differences in performance between no choice and reward choice are accounted for by the extent that individuals receive outcome feedback. The study by Caza et al. (2015) also compares cash, donations to charity, credit points and other types of rewards in an online survey and experiment.[5] The authors show a positive performance effect of reward choice only when people find the reward choice attractive (Caza et al., 2015). Finally, the reward choice in Bareket-Bojmel et al. (2017) consists of choosing between a cash and a tangible reward (cash versus a family pizza meal voucher). In their field experiment, the authors find a positive performance effect when administering a reward choice to employees based on cash versus tangibles (Bareket-Bojmel et al., 2017).

[4] We argue that participants in our study make a reward choice rather than participating in their compensation design. We follow human resource management scholars in arguing that a reward choice is the flexible component included in an employees' total compensation package (Vidal-Salazar et al., 2016).

[5] The authors constructed a measure of reward choice by asking respondents to describe the two most important rewards received from their employer, along with the perceived degree of choice, satisfaction with the degree of choice, and the quality of the available choices (Caza et al., 2015). These six items were then combined into a reward choice measure. In a follow-up experimental study, the authors manipulated reward choice at three levels. More specifically, the levels of reward choice differed in terms of attractiveness of the options included. The options included consisted of cash, credit points, donations to a charity of choice, fast food coupons and donations to the Human Society (Caza et al., 2015).

The reward choices in all the above studies have one common feature in that the reward choice is a choice among rewards that are stored in distinct mental accounts. Thaler's (1999) mental accounting theory posits that distinct types of rewards are stored into different categories, i.e. mental accounts. Evidence indeed suggests that individuals store cash rewards, for instance, into a salary account and tangible rewards into an entertainment account (Choi & Presslee, 2022). These differences in mental account storage, can in turn drive performance effects (Choi & Presslee, 2022; Kelly et al., 2017). While recent evidence indicates that even tangible rewards (i.e. hedonic or utilitarian) can be stored in distinct mental accounts (Mitchell et al., 2021) our study focuses on tangible rewards that are hedonic in nature, therefore keeping the mental account constant. This way, our setting allows us to examine the effect of a mere reward choice.

1.2.2 Hypothesis development

In this study, we argue that a non-task contingent reward choice can incentivize performance in cognitively effortful tasks in gift-exchange settings. To develop our hypotheses, we rely on the basic economic cost-benefit analysis to conjecture an interaction effect of *reward type* and *task type* in terms of cognitive effort, on task performance.

1.2.2.1 The subjective valuation of reward choice

The extent to which individuals will expend cognitive effort in demanding tasks, depends on whether the benefits of such expense will outweigh the costs (Kool & Botvinick, 2018). Exerting cognitive effort is costly (Kool & Botvinick, 2018), and prior findings show that such effort is hard to motivate with financial benefits (Bonner et al., 2000). Hence, we expect that other benefits, like a tangible reward choice, might outweigh the costs of exerting cognitive effort.

Following Mitchell et al. (2021), we argue that tangible rewards worth the same amount will be subjectively valued differently as they are compared to different reference points. Having established that rewards in the choice menu of the present study are all hedonic in nature, we further expect that these rewards will be subjectively valued depending on whether they are chosen or not. We argue that individuals offered the choice of a reward, will subjectively value their tangible reward higher than when the reward is simply allocated. This is because the subjective

value of rewards can be adapted based on the availability of alternative rewards (Otto & Vassena, 2021). In other words, when individuals experience alternatives, they might update the subjective value of their reward of preference. Indeed, individuals derive utility from choosing an option that matches their preferences (Botti & Iyengar, 2004; Givi & Galak, 2017; Oehlmann, Meyerhoff, Mariel, & Weller, 2017). This increased utility in turn, affects satisfaction with the choice made (Iyengar & Lepper, 2000) which in turn can affect how individuals subjectively value the reward choice. Additionally, research on gift-exchange in the workplace, has also shown that workers derive preference matching utility from receiving a choice between a cash or tangible gift, and that they reciprocate by exerting higher effort (Kube et al. 2012). Likewise, a recent study by Lourenço (2020) finds that employees respond positively towards incentives that match their preferences, by increasing their performance.

We argue that the difference in subjective valuation of a chosen versus an allocated tangible reward will in turn affect cognitive effort exertion. More specifically, neuroscience research has established that the prospect of a reward results in greater frontoparietal brain activity, which is the brain region activated during cognitive control (Parro, Dixon, & Christoff, 2018). Moreover, the reciprocal relationship between the prospect of rewards and cognitive effort depends on the value of these rewards (Etzel, Cole, Zacks, Kay, & Braver, 2016; Hall-McMaster, Muhle-Karbe, Myers, & Stokes, 2019; Otto & Vassena, 2021; Westbrook & Braver, 2015). As such, neuroscientists find that the prospect of a higher valued reward increases cognitive effort and accompanied performance improvements more than the prospect of a lower-valued reward (Etzel et al., 2016; Hall-McMaster et al., 2019; Otto & Vassena, 2021).

Turning to the cost-benefit analysis of exerting cognitive effort (Kool & Botvinick, 2018), we reason that individuals will exert cognitive effort according to its subjective costs and the subjective value of the benefit of doing so. In the current context, participants engage in a similar task, which differs to the extent of cognitive effort required to perform the task. Cognitive effort, such as task switching (e.g.. the Stroop task), or taking into account multiple task rules (Dreisbach, 2012), are found to impose subjective costs (Kool, McGuire, Rosen, & Botvinick, 2010). Thus, in the absence of rewards, cognitive effort and performance are found to be

lower on the more demanding versions of the task (Arber et al., 2017; Baumeister et al., 1998; Tice et al., 2007). In the setting of this paper, individuals receive a non-task contingent tangible reward. Based on the theory discussed above, we expect that individuals choosing a tangible reward will subjectively value the reward higher than individuals who are allocated a tangible reward. The higher subjective value of the outcome, can in turn increase the attractiveness of exerting (cognitive) effort.

We thus posit that reward choice will lead to greater effort exertion compared to no reward choice, in demanding tasks. When individuals are offered a reward choice, they derive utility in the form of preference matching which in turn leads individuals to subjectively value the tangible reward higher, than individuals who did not get offered a reward choice. The subjective value of the outcome can in turn cover the subjective costs of exerting effort, because the benefit from doing so increases. Indeed, prior research finds that individuals exert greater cognitive effort given a higher reward, and that exerting cognitive effort might even add value to the overall outcome in such cases (Toro-Serey, Kane, & McGuire, 2021). Likewise, scholars find that benefits in the form of gifts can increase profits when tasks become more demanding (Hesford, Mangin, & Pizzini, 2020), and can increase the performance of health workers whose nature of their jobs is generally more demanding (Brock, Lange, & Leonard, 2018). We therefore argue that offering reward choice affects individuals' motivation to expend more effort compared to no reward choice in a demanding task requiring higher cognitive effort.

In simpler tasks, the beneficial effect of such reward choice may be limited as individuals might be more motivated than in more demanding tasks because the subjective cost of effort is relatively low. As such, Prendergast (1999) argues that rewards are incentivizing for simple jobs. However, a meta-analysis by Deci, Koestner, & Ryan (1999) shows that rewards have a small beneficial effect for tasks that might be experienced as "boring". That is, the evidence available shows that individuals maintain motivation in simple and boring tasks when rewards are being offered. The simple task in this study, detecting letters in text boxes, is assumed to be boring for individuals. Therefore we argue that while reward choice can be motivating, the effect of it might be less pronounced

in easier/boring tasks, given that people are less likely to lose their motivation in simpler tasks.

In sum, theory posits that individuals assign a higher subjective value to a reward chosen than when it is simply allocated to them. As such, individuals seek to maximize the subjective value of their reward by choosing one that closely matches their preferences. A priori, individuals require to exert more effort in tasks that are cognitively more demanding. Consequently, we predict that the effect of reward choice (relative to no choice) on the motivation to expend effort is higher in a more demanding task relative to the easier task. Accordingly, we state our first hypothesis as follows:

Hypothesis 1: *Reward type and task type interact such that reward choice in a cognitively demanding task increases performance more relative to when no reward choice is present.*

From a practical point of view, managers want to keep motivation and performance high across all types of tasks. Therefore, management control systems might provide a solution to motivating performance in more demanding tasks. More specifically, offering a reward choice can incentivize higher performance on demanding tasks because of the heightened subjective value individuals assign this reward. There is, however, considerable tension with regard to this hypothesis as studies in social psychology argue that choice may have no, or even produce a negative effect on performance (see Patall et al. 2008 for a review). Theorists reason that the act of making choices can be effortful, which can result in a state of ego-depletion (Baumeister et al., 1998; Baumeister & Vohs, 2007; Tice et al., 2007; Vohs et al., 2008). In our setting, the reward choice precedes task execution. Given that choices - when considered to be effortful - may reduce cognitive abilities, the question arises whether reward choice might mitigate the positive effects that we described earlier in a demanding task. Nevertheless, there is sufficient evidence in neuroscience that shows that the prospect of a reward can boost motivation to expend effort in these more demanding tasks (Etzel et al., 2016; Hall-McMaster et al., 2019). Therefore we conjecture that reward choice differs in nature to the choices studied in social psychology. More specifically, Iyengar and Lepper (2000) find that offering extensive choice of certain products (i.e. 24 to 30 different

flavors of jam and chocolate) reduces the likelihood of purchasing the product. Likewise, the authors find that students writing an essay performed worse when they chose their topic from a 30-number list compared to a 6-number list (Iyengar & Lepper, 2000). However, the choices in the study by Iyengar and Lepper (2000) are different in nature to the reward choice in this study. That is, the reward choice in this study contains options from three different categories (i.e. experiences, gift cards and merchandise), whereas the former constitutes a choice of a single category (i.e. a product, or an essay topic) where the options constitute differences in attributes (e.g. flavor). Alternatively, prior marketing research finds that choosing from larger assortments, containing both hedonic and utilitarian options, increases choice difficulty and leads consumers to choose for the utilitarian options more than the hedonic options because they are easier to justify (Sela et al., 2009). Nevertheless, the reward choice in this setting contains only hedonic options, so we do not expect these prior findings to hold.

1.3 Experimental method and design

1.3.1 Participants

We recruited 102 participants from a large European university to take part in a compensated laboratory experiment. In addition to one course credit for a management accounting course, participants received a tangible reward, with a value of approximately €10[6], as compensation. Participants signed up for one of twelve experimental sessions, which lasted on average 33 minutes. In each experimental session, one of the four experimental conditions was run. The condition conducted in each session was randomly pre-determined. The participant sample consisted of 69.61% males, with an average age of 22.14 years (SD= 1.39) and a mean work experience of 12.67 months (SD= 15.67).

[6] We follow Jeffrey (2009) and Shaffer and Arkes (2009) by including tangible rewards that have approximately the same retail value.

1.3.2 Design and experimental task

1.3.2.1 Independent variables

The experiment consisted of a 2×2 between-subjects design. The first independent variable *reward type* was related to the participants' compensation. Each participant received a non-task-contingent tangible reward for the experiment. In the *no choice* condition, participants were told that their employer would choose a tangible reward as their compensation, which was randomly drawn from the available options and shown to the participants in the next window. That is, participants in the *no choice* condition only saw the reward they were assigned. In the *reward choice* condition, participants were informed that their employer offered them a choice between 20 tangible rewards to compensate them for their job. Each of the 20 tangible rewards was presented on a separate screen in a random order to avoid order effects, with the final screen showing an overview of all the tangibles where participants could make their final choice.

The second independent variable, *task type* was manipulated by employing two versions of the same letter detection task (Baumeister et al., 1998). The versions of the task differed in the extent to which they required cognitive resources to solve the task. Participants in the *simple task* condition were informed that they worked on the task by crossing out all instances of the letter "e" in the given text boxes. Participants in the *demanding task* condition were informed about the same task, and had to take into account an extra rule while completing the task (i.e. crossing out all letters "e" in the text boxes, but not if there was a vowel adjacent to the "e" or one letter away) (Baumeister et al., 1998; Tice et al., 2007).[7] Each participant could perform the task for a total of ten independent rounds. [8] They were, however, not made aware how many rounds they would have to solve, in order to prevent end-of-task gaming

[7] In previous research, cognitive effortful tasks are typically operationalized as tasks where participants have to override responses (Tice et al., 2007). In this study *task type* was manipulated by including tasks where participants need to control their cognitive resources (i.e. attention, working memory), effort and persistence when facing difficulty or failure, control of impulses etc. (Davis & Leo, 2012).

[8] A pre-test with 13 participants showed that per text box, participants spent 6.18 minutes on a text box, on average. The text boxes were sourced from (Myers et al., 2018) and https://www.rd.com/true-stories/inspiring/100-word-stories/.

(Farrell, Kadous, & Towry, 2008). Their instructions specified that they could work on the task and could freely decide when they wanted to stop with a "stop task button". This design feature allows us to assess, next to task performance, participants' level of task persistence. Namely, prior studies report that overriding impulses (i.e. making use of the limited source of self-control) results in reduced persistence in more cognitively demanding tasks (Schmeichel & Zell, 2007). In addition, poor self-control has been operationalized by quitting sooner on a difficult task (Baumeister et al., 1998; Gailliot et al., 2007), and consequently as a measure of task persistence.

1.3.2.2 The construct of reward choice

The *reward choice* for this study included tangible rewards that belong to three different categories. These tangible rewards were carefully pre-tested. A paper-and-pencil survey that measured the attractiveness of 28 tangible rewards was completed by 280 respondents during a bachelor course in banking and finance at a large European university. Respondents who completed this survey were not eligible to participate in the laboratory experiment. Respondents were asked to rate the attractiveness of 28 tangible rewards (see Appendix 1.1) on a 7 point Likert scale ranging from "very strongly disagree" to "very strongly agree". According to the literature, the included tangible rewards covered the three most preferred categories of rewards among employees, namely experiences, merchandise and gift cards (Haden, 2017; Michalowicz, 2013; The Incentive Research Foundation, 2015). In order to operationalize our *reward type* independent variable, the top five most preferred and three least preferred (attractive) options were excluded from our main experiment, in order to make the option set more balanced w.r.t. reward attractiveness. The resulting portfolio of tangibles included twenty rewards which were all approximately equal in attractiveness ($M=4.88$, $SD=0.71$, $Min=4.30$, $Max=5.39$). The mean attractiveness of the portfolio of tangibles was significantly above the midpoint of 4 ($t=19.82$, one-tailed $p<0.01$) which suggests an attractive portfolio of tangible rewards for participants. The resulting options included five tangibles from the experiences category, seven from the merchandise category and eight from the gift cards category. The menu of tangible rewards in the *reward choice* condition included 20 options, all valued at

approximately €10.[9] For the *no choice* condition, the reward was a randomly assigned tangible reward (out of the 20 available options) equally valued at €10.

1.3.2.3 Dependent variable

The main dependent variable of interest in this study is task performance. We proxy individual's *effort intensity* by measuring task performance (Choi, Clark, & Presslee, 2019). According to Bonner and Sprinkle (2002) effort intensity is referred to 'the amount of attention an individual devotes to a task or activity during a fixed period of time'(p. 306). The authors similarly argue that effort intensity gives an indication of how hard individuals work. As such, the measure of task performance is constructed by the sum of an individual's i proportion of correctly detected letters per text box, divided by the number of completed text boxes n.

$$Task\ performance_i = \frac{\sum_i^n proportion\ of\ correct\ letter\ detections}{n_i}$$

The proportion of correctly detected letters is calculated by dividing the number of correctly detected letters by the total number of letters to be detected, for each given text box. The measure is thus bounded between zero and one, with higher values indicating a higher average task performance.

1.3.3 Procedures

The experimental materials informed participants that they had to complete a task after they had taken notice of the reward that they would receive for participating the experiment. Before learning about the specific task, participants in each of the reward conditions are presented with their reward. In the reward choice condition, they could choose

[9] Next to no choice and extensive reward choice, we ran a third condition. The limited reward choice condition included three options. Both reward choices represented all three categories (i.e. experiences, merchandise and gift cards). While the extensive reward choice included more options per category, the nature of the limited reward choice was the same. That is, increasing the choice set from limited to extensive choice reflects a greater number of options per category. We analyze this third condition in our supplemental analyses.

among 20 rewards, whereas in the no choice condition one of the 20 options was randomly assigned to them (i.e. they did not have knowledge on the other 19 options).[10] The reason for having the subjects choose their reward before working on the task is that the knowledge about the received reward remains constant across all conditions. Also, the study by Heninger et al. (2019) offers a reward choice upfront, which provides us with practical evidence. Additionally, we need participants to choose upfront for our theory, which states that the subjective value of the chosen reward will be higher than the subjective value of an assigned reward (due to preference matching). Participants were informed that they would receive this reward for participating in the experiment, making it non-contingent on task performance (Ryan, Mims, & Koestner, 1983). Participants thus all received their reward regardless of their performance in the experiment. This design choice ensures that the motivating effect of reward choice can be measured. This way, we minimize the extent to which our results can be attributed to confounding effects. Previous research in management accounting also makes use of this specific design choice (Hales et al., 2015). After having learned of their compensations, participants received instructions about the experimental task. Within the instructions, it was clearly described to the participants that they could stop working on the task at any time they wanted (Beckers, Cardinaels, Dierynck, & Yin, 2018; Hales et al., 2015). Each screen was provided with a button to stop the task.

Finally, participants completed a post-experimental questionnaire that measured personality traits such as personal growth initiative (Robitschek et al., 2012). Moreover, questions about the final tangible reward and the letter detection task were also included. At the very end of the post-experimental questionnaire participants were also asked to disclose their demographic information and their contact details for pay out of their reward.

[10] Rewards were paid during the month succeeding the experiment.

1.4 Results

1.4.1 Descriptive statistics, manipulation and randomization checks

Descriptive statistics for the dependent variable used in the hypothesis test are presented in Table 1.1, panel A, and presented graphically in Figure 1.1. Table 1.1, Panel B presents the descriptive statistics for a set of alternative dependent variables for our hypothesis test. Descriptive statistics for variables used in additional analyses are presented in Table 1, panel C, i.e. the items questioned in the post-experimental questionnaire (PEQ). The correlations between these variables are presented in Table 1.2.

As can be seen from Table 1.1, the task performance rates in the *reward choice* condition for the *simpler task* (M=0.84, SD=0.20) and the *demanding task* (M=0.80, SD=0.16) are fairly similar. Figure 1.1 also shows these means graphically. Nevertheless, the difference in task performance for both conditions is significant (Wilcoxon rank-sum test Z=2.17; n_1=27; n_2=25; p<0.03 (two-tailed)). Similarly, there is a clear difference in task performance rate for the two task versions in the *no choice* condition (*demanding task*: M=0.68, SD= 0.20; and *simple task*: M=0.88, SD=0.12). This difference is significantly different from zero (Wilcoxon rank-sum test Z=4.04; n_1=26; n_2=24; p<0.001 (two-tailed)). This result is in line with earlier findings (Arber et al., 2017; Baumeister et al., 1998). In addition, within the *demanding task, reward choice* results in a significant higher task performance rate (M=0.80, SD=0.16) as compared to *no choice* (M=0.69, SD=0.20) (Wilcoxon rank-sum test Z=-2.43; n_1=26; n_2=25; p<0.02 (two-tailed)). Finally, within the *simple task*, *reward choice* results in a lower task performance rate (M=0.84, SD=0.20) as compared to *no choice* (M=0.88, SD=0.12), however this difference is small and not significant (Wilcoxon rank-sum test Z=0.51; n_1=24; n_2=27; p>0.61 (two-tailed)).

Table 1.1: Descriptive statistics

		Reward Choice		No Choice	
		Simple Task	**Demanding Task**	**Simple Task**	**Demanding Task**
Panel A: Dependent variables					
Task performance		0,842	0,798	0,878	0,683
		(0,201)	(0,161)	(0,121)	(0,196)
Panel B: Alternative variables					
	Quitters	0,185	0,440	0,333	0,231
Effort		(0,396)	(0,507)	(0,482)	(0,430)
Duration	Completed	9,222	8,440	8,792	8,654
	Textboxes	(2,118)	(2,567)	(2,064)	(2,637)
Overall performance		0,805	0,683	0,782	0,595
		(0,229)	(0,257)	(0,218)	(0,260)
Panel C: PEQ items					
Personal Growth[a]		5,044	4,936	5,008	5,146
		(0,937)	(0,971)	(0,839)	(1,213)
Preference Matching[b]		5.296	5.680	3.615	4.048
		(0.963)	(0.789)	(1.361)	(1.564)
Number of participants		27	25	24	26

Panel A presents the mean, (standard deviation) for the dependent variable for each condition. Task performance is measured as the percentage of correct letter detections, divided by the number of text boxes.

Panel B presents the mean, (standard deviation) for the alternative variables for each condition. Quitters is a binary variable measured as the number of participants indicating to quit on the 10th round or earlier. Completed Textboxes is a discrete variable measured as the round in which participants indicate to quit. Overall Performance is measured by the percentage of correct letter detections over all rounds.

Panel C presents the means, (standard deviation) for the post-experimental questionnaire items, all measured on a 1-7 point Likert scale.

[a] Personal growth represents participants' degree of willingness to grow. The scale is constructed using 5 items (Robitschek et al., 2012) that measures the extent to which participants agree with statements their plans about personal growth on a 7 point Likert scale (1=strongly disagree, 4=neutral, 7=strongly agree). Confirmatory factor analysis with principal component factoring estimation results in a one factor solution (based on eigenvalue > 1). The scale has good internal consistency with a Chronbach's alpha of 0.86. For ease of interpretation, we report mean scores based on combining the equally-weighted means of the 4 items included.

[b] Preference matching represents participants' assessment of how they feel that their reward matches their preferences using a 7 point Likert scale (1=strongly disagree, 4=neutral, 7=strongly agree). The scale was constructed by including the following items "I am satisfied with the tangible reward that I received for the task."; "The tangible reward that I received for the task, matches my preferences."; "I strongly identified with the tangible reward I will receive for this study." And "I think the tangible reward I will receive for this study is attractive."; all measured on a 7 point Likert scale. Confirmatory factor analysis with principal component factoring estimation results in a one factor solution (based on eigenvalue > 1). The resulting scale had good

internal consistency α=0.91. For ease of interpretation, we report mean scores based on combining the equally-weighted means of the 4 items included.

Table 1.2: Pearson correlations

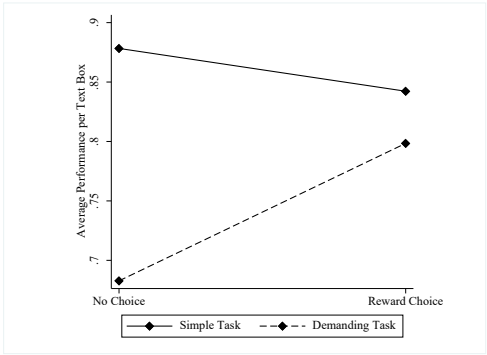
	1.	2.	3.	4.	5.	6.
1. Task Performance	1					
2. Quitters	-0.151**	1				
3. Completed Textboxes	0.253***	-0.846***	1			
4. Overall Performance	0.771***	-0.648***	0.772***	1		
5. Preference matching	0.088	0.115	0.074	-0.081	1	
6. Personal growth	0.015	-0.135*	0.183**	0.086	-0.100	1

*, **, *** denote significance at the 10%, 5%, and 1% level (two-tailed), respectively. Correlations are based on the principal component factor scores.

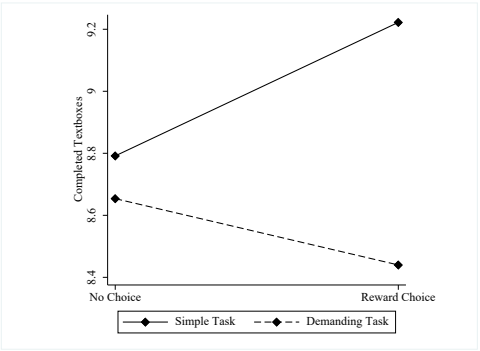
Figure 1.1: Graphical representation of dependent variables elicited in the experiment

Graphical depictions for the average performance per text box (Panel A), the round in which participants quit (Panel B), and the proportion of participants quitting (Panel C), by condition.

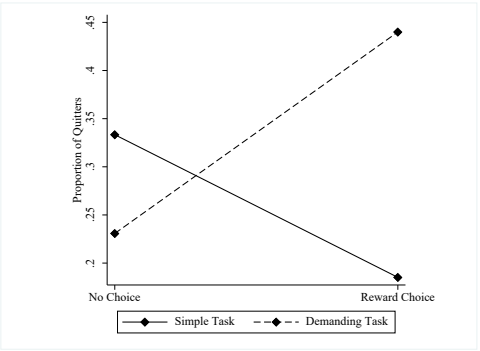
Panel A:



Panel B:



Panel C:



Participants work on the experimental task, and they can indicate whether or not they want to continue with the next round (i.e. text box).
Panel A shows the average percentage of correct letter detections per text box, by condition.
Panel B shows the round in which participants quit (completed textboxes), by condition.
Panel C shows the proportion of participants who indicated to quit before ending the last round of the experiment, by condition.

To check the effectiveness of our *reward type* manipulation, we asked participants in the PEQ whether their tangible reward earned for this experiment was chosen by themselves or by their employer. From the participants in the *reward choice* condition, 92% of them indicated that they chose their reward themselves. Alternatively, all participants in the *no choice* condition indicated that their reward was chosen by their employer. Thus, almost all participants were aware of who had chosen their reward, suggesting that our reward type manipulation was successful.[11]

To assess whether the *task type* manipulation was efficient, we asked participants in the PEQ whether or not they had to take into account an extra rule when completing the task, besides highlighting every letter "e" (Baumeister et al. 1998, Tice et al. 2007). Again, all participants in the *demanding task* condition reported that they had to take into account an extra rule while completing the task (100%). Furthermore, 93.5% of the participants in the *simpler task* condition understood that they did not have to take into account an extra rule for completion of the task. Additionally, we asked participants to indicate the extent to which they agreed to the statement "the detection task was challenging" on a 7-point Likert scale (1=strongly disagree, 7=strongly agree). Average responses were higher for participants in the demanding task conditions (M=3.37, SD=0.22) than for participants in the simpler task condition (M=2.90, SD=0.22), and this difference was marginally significant (t_{100}=-1.53, one-tailed p-value=0.07).[12] This suggests that our *task type* manipulation was successful.

Finally, we assess whether randomization was successful. We conducted 12 sessions, in which one experimental condition was

[11] In order to check whether our manipulation was successful for our original, nested conditions, the *limited* reward choice condition, containing three options and the 20 options in the *extensive* reward choice condition, we compared the mean time spend on making the choice. The mean time for *limited* (M=25.13 seconds, SD=11.76 seconds) and mean time for *extensive* choice (M=101.81 seconds, SD=35.92) was significantly different from zero (t(91)=-13.11, p-value<0.001, two-tailed), which suggest successful manipulation.

[12] When considering the observations from the limited choice condition, this result still holds. Likewise, the average response on this PEQ item was higher for participants in the demanding task condition (M=3.43, SD=0.18) than for participants in the simpler task condition (M=2.71, SD=0.16), and this difference was significantly different from zero (t_{151}=-2.98, two-tailed p-value=0.004).

conducted. This was done to ensure balanced cell sizes. The conditions were randomly pre-determined. We do not find significant effects of our manipulations for the variables age, gender, language proficiency and English reading habits (untabulated, all two-tailed p-values > 0.26). However, when considering work experience as dependent variable, we find a significant effect of our task type manipulation (two-tailed p-value = 0.06). That is, participants in the more demanding task type have significantly more working experience than participants in the simpler version of the task.[13]

1.4.2 Hypothesis test

Recall that our hypothesis predicts an interaction effect of *reward type* and *task type*, such that we expect *reward choice* to lead to higher task performance when the task becomes more cognitively demanding. To test our hypothesis formally, an analysis of variance (ANOVA) is conducted. Table 1.3, Panel A shows the results of the ANOVA analysis in which task performance (i.e. the proportion of correct letter detections per text box solved) is the dependent variable. We find evidence of a significant interaction effect of *reward type* and *task type* (F=4.84; *p*=0.03 (two-tailed)). This result provides evidence for our hypothesis, which posits that the effect of *reward choice* compared to *no choice* on task performance depends on the level of cognitive effort required to complete a task. Additionally, the simple effects (Table 1.3, Panel B) show that in the *demanding task*, *reward choice* works better than simple rewards, while this is not the case for the *simple task*.

[13] Results for our hypothesis are inferentially the same after controlling for participant work experience.

Table 1.3: Hypothesis test

Panel A: ANOVA on task performance				
	df	**M.S.**	**F**	**p-value**
Task	1	0.365	12.06	0.001***
Reward Type	1	0.040	1.34	0.251
Task x Reward Type	1	0.146	4.84	0.030**
Residual	98	0.030		
Panel B: Follow-up simple effects				
		df	**F**	**p-value**
Effect of $Task$ in the No Choice condition		1	15.79	0.000***
Effect of $Task$ in the Reward Choice condition		1	0.83	0.366
Effect of $Reward\ Type$ in the Demanding task		1	5.64	0.020***
Effect of $Reward\ Type$ in the Simple task		1	0.55	0.462

***, **, * Indicate p-values at $p < 0.01$, $p < 0.05$, and $p < 0.10$, respectively.
All p-values are reported two-tailed.
Adjusted R-squared of the ANOVA analysis is 0.132.

1.4.2.1 Preference matching effect

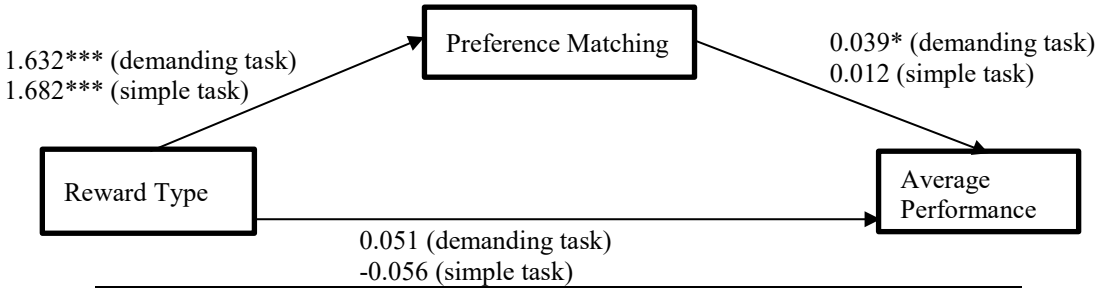
Our theory underlying the effect of reward choice on average performance posits that individuals derive utility from matching the reward to their preferences, which heightens the subjective value of the tangible reward, and consequently makes exerting cognitive effort less costly. As such, we expect that preference matching will mediate the relationship between reward choice and average performance, in demanding tasks. To measure preference matching, we ask participants the following questions: (1) "I am satisfied with the tangible reward that I received for the task", (2) "The tangible reward that I received for the task, matches my preferences", (3) I strongly identified with the tangible reward which I will receive for this study", and (4) "I think the tangible reward I will receive for this study is attractive". Participants were asked to indicate the extent to which they agreed with the above statements, using a 7 point Likert scale (1=strongly disagree, 7=strongly agree). We conducted a confirmatory factor analysis using principal component factoring, which resulted in a one factor solution (eigenvalue=2.82). All of the factor loadings were greater than 0.70, indicating that all four items load high on the same factor. We constructed the measure for preference

matching by averaging the responses to the above four items. The resulting scale was showed high internal validity (α=0.91).

We test whether preference matching mediates the relationship between *reward type* and task performance following Hayes' (2018) Model 4. The mediation model is graphically depicted in Figure 1.2, panel A. The results indeed show that preference matching mediates the relation between *reward type* and task performance in cognitively demanding tasks. More specifically, we find that *reward choice* leads to higher preference matching (1.68; 90% CI [1.08, 2.18]), which in turn causes higher task performance (0.04; 90% CI [0.00, 0.08]). However, the indirect effect of *reward choice* on task performance through preference matching is close to conventional significance levels, but fails to be significant (0.06; 60% CI [-0.02, 0.15]), see Figure 1.2, Panel B. Alternatively, we do not find evidence for this mediation in the simpler version of the task, see Figure 1.2, Panel C. In sum, these findings provide support for our theory. We argue that individuals in cognitive demanding tasks derive preference matching utility from a reward choice, which in turn increase the subjective value of the reward and consequently (cognitive) task performance. This is in line with prior findings from neuroscience (Etzel et al., 2016; Hall-McMaster et al., 2019).

Figure 1.2: Mediation analysis

Panel A: Mediaton Model



Panel B: Bootstrap Results – demanding task

	Effect	**Bootstrapped SE**	**Lower 90% CI**	**Upper 90% CI**
Indirect effect	0.064	0.041	-0.004	0.132
Total effect	0.116	0.0478	0.037	0.194

Panel C: Bootstrap Results – simple task

	Effect	**Bootstrapped SE**	**Lower 90% CI**	**Upper 90% CI**
Indirect effect	0.019	0.028	-0.026	0.066
Total effect	-0.036	0.048	-0.115	0.043

Panel A presents the proposed mediation model graphically for participants task type conditions (simple version versus more demanding version), showing the direct effects (standardized coefficients) of the presented variables. Reward Type is defined as a binary variable indicating 1 when participants are given a tangible reward choice and 0 otherwise. Preference matching is measured as response to "The tangible reward that I received for the task, matches my preferences", which is a statement measured on a 7-point Likert scale (1=strongly disagree, and 7=strongly agree). Average performance is measured as the average correct letter detections over participants' solved text boxes.
Panel B presents the bootstrapped estimations following Hayes (2018) with bias-corrected confidence intervals, for participants in the demanding task version.
Panel C presents the bootstrapped estimations following Hayes (2018) with bias-corrected confidence intervals, for participants in the simple task version.
***, **, * Indicate two-tailed p-values at $p < 0.01$, $p < 0.05$, and $p < 0.10$, respectively.

1.4.3 Supplementary analyses

1.4.3.1 Effect of reward choice on effort duration and overall performance

While we do find a positive performance effect of *reward choice* in the *demanding task* (see hypothesis test), Figure 1.1 Panel C hints at a potential downside of offering reward choices in more demanding tasks.

As such, we also consider an alternative variable *effort duration*. The latent construct of effort duration, measures how long a person works on a particular task. Effort duration is typically defined as the extent "the length of time an individual devotes cognitive and physical resources to a particular task or activity" (Bonner & Sprinkle, 2002, p. 306). Similar as Anand, Webb and Wong (2019) we proxy *effort duration* by two measures. Our first measure of effort duration is *quitters,* defined as the proportion of participants that indicated to end working on their experimental task before finishing the tenth and final round (i.e. text box) of the experiment. By analyzing the proportion of *quitters* we proxy for the degree of participants' motivation (Deci & Ryan, 1985; Hales et al., 2015). If participants quit, we can say that their motivation to work on the task is lower (Anand, Webb and Wong 2018). More specifically, *quitters* is a binary variable that takes on the value of 1 when the participant quitted before ending all rounds of the experimental task, and 0 otherwise. Those participants indicating that they wanted to continue after the tenth round were not labelled as quitters, those who did quit in round ten or in earlier rounds were labeled as quitters.[14] In all experimental conditions, there is a proportion of the participants that quit before the actual end of the experiment. From Figure 1.1, Panel C it can be seen that participants tend to quit most (44%) when presented with a *reward choice* in the *demanding task*. On the contrary, looking at the quitting rate of only 19%, we can conclude that subjects were most motivated or persistent to keep on working on the task when confronted with a *reward choice* in the *simple task*. The difference between these quitting rates are significant (Wilcoxon rank-sum test Z=-1.97; n_1=27; n_2=25; p=0.05 (two-tailed)) suggesting that individuals quit more when given a *reward choice* and confronted with a cognitive demanding task type. We test this formally by conducting a logistic regression analysis on quitting (see Table 1.4, Model 1). The estimated coefficients indicate a positive significant interaction between $Reward\ Type \times Task\ Type$ (β=1.75; p=0.05 (two-tailed)).

[14] This is confirmed by the data, self-reported motivation (measured in the post-experimental questionnaire) negatively correlates with our dependent variable. The correlation coefficient is -0.1776 (p-value = 0.03) is significant at the 5% level.

Table 1.4: Results of regressions

Dependent variable:	**Model 1: DV=Quitting**	**Model 2: DV=Overall Performance**
Demanding Task	-0.511	-0.187***
	(0.630)	(0.068)
Reward Choice	-0.788	0.022
	(0.658)	(0.068)
Demanding Task x Reward Choice	1.751*	0.066
	(0.901)	(0.096)
Constant	-0.693	0.782***
	(0.433)	(0.049)
R^2	0.039	0.085
N	102	102

***, **, * Indicate p-values at $p < 0.01$, $p < 0.05$, and $p < 0.10$, respectively.
Model 1 presents the results of a logistic regression on quitters. All p-values are reported two-tailed. R^2 represents the pseudo R^2 computed using "logit" command in STATA.
Model 2 presents the results of an OLS regression on Overall Performance. All p-values are reported two-tailed. R^2 represents the adjusted R^2, computed using "regress" command in STATA.
Variable definitions:
Reward Type is a binary variable that represents the experimental conditions and equals one if a reward choice was offered to the participant, and zero otherwise. Task is also a binary variable indicating one if the experimental task was the demanding task, and zero otherwise. Quitters is a binary variable measured when a participant indicates that (s)he wishes to quit before finishing the tenth round in the experiment. Effort Intensity is measured as the percentage of correct letter detections over all ten rounds (i.e. text boxes).

Our second measure of *effort duration* indicates the final number of text boxes completed. In Figure 1.1, Panel B we graph completed textboxes across conditions. However, we find no significant interaction effect. Likewise, the follow-up simple effect of the ANOVA analysis (Table 1.5) show that there is no significant difference in when participants quit in the *reward choice* condition ($p > 0.24$ two-tailed), nor a significant effect of *reward type* in the *demanding task* ($p > 0.75$ two-tailed). While we find no significant effects, the findings point to a suggestion that individuals quit earlier in our experiment when offered no choice in rewards than when a reward choice is offered, in a more demanding task.

Table 1.5: ANOVA on effort duration

Panel A: ANOVA on completed textboxes				
	df	**M.S.**	**F**	**p-value**
Task	1	5.386	0.97	0.328
Reward Type	1	0.298	0.05	0.818
Task x Reward Type	1	2.642	0.47	0.493
Residual	98	6.548		
Panel B: Follow-up simple effects				
		df	**F**	**p-value**
Effect of $Task$ in the No Choice condition		1	0.04	0.837
Effect of $Task$ in the Reward Choice condition		1	1.42	0.236
Effect of $Reward\ Type$ in the Demanding task		1	0.10	0.747
Effect of $Reward\ Type$ in the Simple task		1	0.42	0.517

***, **, * Indicate p-values at $p < 0.01$, $p < 0.05$, and $p < 0.10$, respectively.
All p-values are reported two-tailed.
Adjusted R-squared of the ANOVA analysis is -0.015.

Admittedly, the effect of *reward type* and *task type* is two-fold. One the one hand, offering a reward choice increases *task performance* in more demanding tasks while, on the other hand simultaneously increasing the extent to which participants quit. However, this finding should be interpreted with care, as the results also indicate that quitting materializes later in more demanding tasks when offered a reward choice, compared to when no reward choice is present. Similarly, we test the effects of our manipulations on *overall performance,* which is a continuous variable bounded between 1 and 0. We measure *overall performance* as the average performance of participants over all ten text boxes. As individuals have the option to quit, we consider the 10^{th} text box as the period limit of the task. Thus, the measure of *overall performance* takes into account both average performance on performed text boxes and quitting behavior. The results of a linear regression analysis on *overall performance* indicates that the interaction of our manipulations do not affect overall performance (Table 1.4, Model 2). The results shown in. Table 1.4 thus suggest that quitting might be a side effect of offering reward choice in more demanding tasks.

1.4.3.2 Effect of personal growth

Prior research finds that certain types of individuals are more likely to enjoy having a large number of options to choose from. More specifically, more ambitious workers have been found to be more

energized by larger choice sets (Chua & Iyengar, 2006).[15] Therefore, we examine whether *reward type* in our study affects different types of individuals. We make use of the personal growth initiative scale by Robitschek (2012), focusing on the construct of planfulness. Table 1.6, Panel C shows the mean scores for the personal growth scale, per condition.[16] Participants scoring above the median on this scale, are assigned to the high personal growth subsample and the remaining participants are assigned to the low personal growth subsample. Personal growth initiative moderates the relationship between our manipulations and average performance (see Table 1.6). The three-way interaction term between *task type*, *reward type* and *personal growth*, is significant. As such, we find that offering reward choice to high personal growth individuals increases average performance more compared to low personal growth individuals, in cognitively demanding tasks (p=0.04 two-tailed). Table 1.7, panel A presents the results of an ANOVA on a subsample of the participants scoring high on the personal growth scale. The follow up simple effects show that the effect of reward choice is highly significant in demanding tasks for high personal growth individuals (p=0.02 two-tailed). On the contrary, reward choice has no significant effect on average performance for low personal growth individuals (p>0.52 two-tailed, untabulated).

Table 1.6: Three-way ANOVA on personality scale

Panel A: ANOVA on task performance – full sample				
	df	**M.S.**	**F**	**p-value**
Task	1	0.302	10.520	0.002**
Reward Type	1	0.021	0.750	0.399
Personal Growth	1	0.017	0.580	0.450
Task x Reward Type	1	0.138	4.820	0.031**
Task x Reward Type x Personal Growth	3	0.081	2.840	0.042**
Residual	94	0.029		

***, **, * Indicate p-values at p < 0.01, p < 0.05, and p < 0.10, respectively.
All p-values are reported two-tailed.
Adjusted R-squared of the ANOVA analysis is 0.175.

[15] Chua and Iyengar (2006) show this using individual regulatory focus to measure the extent to which individuals are promotion – or prevention - focused (Higgins, 1996).
[16] We find no significant effect of our manipulations on personal growth scores (all two-tailed p-values > 0.54).

Table 1.7: Sample split analyses for high personal growth individuals

Panel A: ANOVA on task performance for high personal growth individuals (subsample)				
	df	**M.S.**	**F**	**p-value**
Task	1	0.325	15.39	0.001***
Reward Type	1	0.078	3.71	0.062*
Task x Reward Type	1	0.053	2.49	0.123
Residual	38	0.021		
Panel B: Follow-up simple effects				
		df	**F**	**p-value**
Effect of $Task$ in the No Choice condition		1	14.51	0.001***
Effect of $Task$ in the Reward Choice condition		1	2.87	0.098*
Effect of $Reward\ Type$ in the Demanding task		1	6.55	0.015**
Effect of $Reward\ Type$ in the Simple task		1	0.06	0.812

***, **, * Indicate p-values at $p < 0.01$, $p < 0.05$, and $p < 0.10$, respectively.
All p-values are reported two-tailed.
Adjusted R-squared of the ANOVA analysis is 0.412.

1.4.3.3 Limited reward choice

The original experiment included a reward choice condition with only three options. The *limited* choice condition included three rewards, one for each category (i.e. experiences, merchandise and gift card). Ten distributions of three reward options were pre-programmed from which one was randomly drawn to be presented to the participants. This design choice was executed in order to achieve comparability to the *extensive* choice condition in terms of attractiveness and to reduce the risk of over representing the same award. As such, there was a high likelihood that most of the rewards from the above set were also represented in the *limited* condition. A total of 51 participants were offered the *limited reward choice*, of which 25 were assigned to the *demanding task* condition, and 26 to the *simple task* condition.

In this supplemental analysis, we show the similarity between this *limited reward choice* condition and the *no choice* condition. Table 1.8 and Table 1.9 depict our analyses on our dependent variables average task performance and quitters, respectively, distinguishing between the *limited reward choice* and *no choice* levels of our *reward type* manipulation. Both tables show no significant effects of our main independent variable of interest (i.e. *reward type*). A possible explanation for these non-results could be that offering a limited choice between three options only, does not allow for preference matching and therefore does

not affect the subjective value of the chosen tangible reward. Moreover, research from social psychology suggests that when a choice does not confer sufficient control to an individual, motivating effects of such choice do not occur (Sullivan-Toole, Richey, & Tricomi, 2017).

Table 1.8: ANOVA on task performance - limited vs. no choice

Panel A: ANOVA on task performance				
	df	**M.S.**	**F**	**p-value**
Task	1	1.139	44.89	0.000***
Reward Type	1	0.022	0.85	0.360
Task x Reward Type	1	0.007	0.29	0.592
Residual	97	0.025		
Panel B: Follow-up simple effects				
		df	**F**	**p-value**
Effect of $Task$ in the No Choice condition		1	18.79	0.000***
Effect of $Task$ in the Limited Choice condition		1	26.47	0.000***
Effect of $Reward\ Type$ in the Demanding task		1	0.07	0.787
Effect of $Reward\ Type$ in the Simple task		1	1.05	0.308

***, **, * Indicate p-values at $p < 0.01$, $p < 0.05$, and $p < 0.10$, respectively.
All p-values are reported two-tailed.
Adjusted R-squared of the ANOVA analysis is 0.303.

Table 1.9: Logistic regression on quitters - limited vs. no choice

Dependent variable:	**Quitters**
Demanding Task	0.511
	(0.636)
Reward Choice	-0.118
	(0.607)
Demanding Task x Reward Choice	0.377
	(0.885)
Constant	-0.693
	(0.433)
R^2	0.006
N	101

***, **, * Indicate p-values at $p < 0.01$, $p < 0.05$, and $p < 0.10$, respectively.
All p-values are reported two-tailed. R^2 represents the pseudo R^2 computed using "logit" command in STATA.
Variable definitions: Reward Type is a binary variable that represents the experimental conditions and equals one if a *limited* reward choice was offered to the participant, and zero otherwise. Task is also a binary variable indicating one if the experimental task was the demanding task, and zero otherwise. Quitters is a binary variable measured when a participant indicates that (s)he wishes to quit before finishing the tenth round in the experiment.

1.5 Conclusion and discussion

This study investigates the effect of reward choice across two settings which differ in their degree of cognitive resources needed to complete the job. The results show that offering individuals a reward choice can motivate cognitive task performance. Nevertheless, this beneficial effect of reward choice comes at a cost as we find that participants are most prone to quit when confronted with an extensive choice in a cognitive costly task environment. These results are in line with earlier findings that option size influences subsequent self-control and therefore performance (Iyengar & Lepper, 2000). However, this side effect should be interpreted with care, as we find that participants quit later in the reward choice condition (compared to those in the no choice condition).

Whereas Bonner et al. (2000) conjecture that incentives are less likely to improve performance in tasks that become more cognitively demanding, we find evidence of an incentive in the form of reward choice that can improve performance in such tasks. We find evidence suggesting that such incentives increase individual productivity. Additionally, we find that reward choice increases performance most for individuals with high personal growth initiative.

Building on prior research examining tangible rewards, our findings suggest that a non-task contingent tangible reward choice can increase (cognitive) task performance. With this study we partly respond to Mitchell and colleagues'(2021) call to investigate the effects of tangible rewards in compensation schemes where pay is not contingent on performance. However, the extent to which tangible reward nature (hedonic versus utilitarian) affects performance in non-contingent pay settings, remains to be studied (Choi & Presslee, 2022; Mitchell et al., 2021).

While scholars argue that the optimal amount of options to be included in a menu of choices is three (Iyengar & Lepper, 2000), we do not find support for this claim in this study. Our findings show that the differences between the no choice condition and reward choice condition containing three options are not significantly different from zero (untabulated). These results suggest that for a reward choice to have an effect on either employee performance or employee motivation, it should be sufficiently large.

Our findings also have practical implications. We provide direct evidence of a two-folded interaction effect of reward choice and task type

on average performance. Our results essentially suggest that offering reward choice can positively affect task performance across both simple and more demanding tasks, as we observe small and insignificant differences in task performance across our simple task condition. Nevertheless, we do observe a potential cost to offering reward choice in more demanding task types which materializes in lower effort duration. This could potentially lead to individuals turning to other activities such as cyber-loafing for example (Koch & Nafziger, 2016). Therefore, firms might carefully consider when to include large menus of rewards as part of employees' compensation packages. That is, employees with jobs that demand substantial cognitive resources, such as health care workers for example, might benefit from choosing their reward more than employees who have a more routinely job. Nevertheless, one might argue that both versions of the letter detection task in this study might be characterized as routine or low-skilled jobs. Examples of routine jobs can be found packaging, assembling in manufacturing. The extent to which these jobs can become more demanding is typically reflected in the extent to which people need an analytical aspect as compared to manual work in their manufacturing or industrial job (De Vries, Gentile, Miroudot, & Wacker, 2020). This can refer to monitoring, measuring and controlling activities. With our study, we speak to these types of jobs. Our findings thus suggest that while individuals who are highly ambitious (scoring high on the personal growth initiative scale by Robitschek et al. (2012)) are effectively incentivized to perform well on more demanding routine jobs (such as monitoring and controlling activities), they might also want grow into more senior and higher skilled jobs. Therefore, future research is needed to study the longer term effects of reward choices in these job types, as our findings might hint to potential losses of offering reward choice from an organizational perspective.

As with any experimental study, this study is subject to limitations which can highlight opportunities for further research. First, while we use a gift-exchange setting in which tangible reward payout is not contingent on task performance to examine the motivating effects of reward choice, prior evidence suggests that the effect of reward type can depend on the type of incentive scheme offered (Mitchell et al., 2021; Presslee et al., 2013). Further research could add to this by examining whether reward choice has different effects in a complete contracting setting like a flexible reward type (Baeten & Verwaeren, 2012; Choi & Presslee, 2022;

Kube, Marchal, & Puppe, 2012). Second, the rewards in this study were paid out to the participants about 4 weeks after the experimental sessions. As previous research shows that immediate versus lagged payout affect behavior differently, future research could look into these effects and examine whether the effects of reward choice persist in a longer term time frame (Becker, Messer, & Wolter, 2013; Heneman, Fisher, & Dixon, 2001; Keh & Lee, 2006). Additionally, recent evidence by Choi and Presslee (2022) indicates that tangible rewards are subjectively valued higher than utilitarian rewards because they are usually novel and unexpected. However, participants in our setting were aware of the tangible reward they would receive for completing the study, before actual task execution. While we opted for this design choice to ensure similar information w.r.t. the reward in all our conditions, we acknowledge that it could have affected our results. Therefore, future research could examine whether our results would still hold if participants only receive the information of a reward choice upfront, without explicitly mentioning the content of such reward choice such that rewards remain novel and unexpected. Third, participants in the experiment were either told that their employer had given them the opportunity to choose a reward or that their employer had chosen one for them. Future research could look into the effect of social distance to the hypothetical employer and the different responses in behavior that this entails (Charness, 2000; Hoffman, McCabe, & Smith, 1996). Moreover, the fact that participants were told that their employer was responsible for choosing the way in which they were compensated, could have induced reciprocity effects. Nevertheless, our study did not explicitly measure for reciprocity, therefore we cannot disentangle whether our reward choice effect was driven by either heightened subjective value (because of preference matching) or reciprocity (Becker et al., 2013; Berg, Dickhaut, & McCabe, 1995; Boosey & Goerg, 2020; Bradler, Dur, Neckermann, & Non, 2016; Bradler & Neckermann, 2019; Fehr & Gachter, 2000). Finally, participants in this study were subject to an uninteresting, boring experimental task. While rewards have been shown to incentivize motivation in such task types (Deci et al., 1999), prior research has established a crowding out effect of rewards on intrinsic motivation (Deci, 1971). Therefore, future research might study whether and how reward choice might undermine intrinsic motivation in interesting or fun task types.

Appendices

Appendix 1.1: Descriptive statistics for attractiveness of pre-tested tangibles rewards

Tangible reward	N	Mean	SD	Min	Max
Visit to a large brewery	280	4.39	1.65	1	7
Visit to Museum	280	3.24	1.37	1	7
Movie ticket	280	5.71	1.17	1	7
Voucher for spa treatment	280	4.85	1.77	1	7
Kayak activity	280	4.82	1.61	1	7
Voucher for an adventure park activity	279	5.23	1.41	1	7
Concert visit	278	5.08	1.43	1	7
University mug	278	3.65	1.64	1	7
University water bottle	278	4.75	1.59	1	7
University notebook	278	3.47	1.63	1	7
University t-shirt	279	4.30	1.58	1	7
A box of chocolates	279	5.39	1.62	1	7
A beer set	279	4.79	1.92	1	7
An Italian gift basket	279	5.32	1.42	1	7
A fresh fruits gift basket	279	4.79	1.44	1	7
A sweets and candy gift basket	279	4.84	1.65	1	7
A “Zalando” gift card	279	5.89	1.39	1	7
An “Asos” gift card	279	5.26	1.67	1	7
An “About you” gift card	279	4.85	1.61	1	7
A “bol.com” gift card	279	6.10	1.06	2	7
A local bookshop gift card	279	5.69	1.36	1	7
Another local bookshop gift card	278	4.78	1.61	1	7
An “Amazon” gift card	279	5.57	1.37	1	7
A gift card for a local café	279	4.52	1.73	1	7
Another gift card for a local café	279	4.40	1.70	1	7
A “Starbucks” gift card	279	4.92	1.75	1	7
A gift card for a burger restaurant	279	5.39	1.43	1	7
A gift card for a local restaurant	279	4.95	1.58	1	7

Chapter 2

Peer evaluations: the effects of system design and outcome transparency on employee effort*

Abstract:

In this paper, we examine the effect of outcome transparency for given control systems as we investigate two types of peer evaluation systems. We draw on self-concept maintenance theory to predict a disordinal interaction between peer evaluation system and outcome transparency. We collect data through an online 2x2 between-subjects experiment where employees work on an image description task, where they allocate effort towards quantity and quality. We manipulate the peer evaluation system as a rating or ranking system and whether or not evaluation outcomes are made transparent to peers. Our results suggest that peer rankings relative to peer ratings seem to mitigate the negative effect of ratings on employee effort found in prior literature (Carpenter et al., 2010) when its outcomes are kept private. Alternatively, we find that peer ratings incentivize employee effort more when peer evaluation information is transparent compared to peer rankings. We find similar results when we correct for the quality of image descriptions. Collectively, our results contribute to a better understanding of peer evaluation systems in practice and how they should be designed to promote employee effort.

* This chapter is co-authored with Eddy Cardinaels and Alexandra Van den Abbeele. We thank Markus Arnold, Martine Cools, Sophie De Winne, Christoph Feichter, Sabra Khajehnejad, Jonathan Gay (discussant), anonymous reviewers and conference participants at the VIII Research Forum on Challenges in Management Accounting 2021, the 11th EIASM conference on performance measurement and management control 2021, and the AAA Management Accounting Section Midyear Meeting 2022.

2.1 Introduction

In certain organizational settings, like crowdsourcing settings, for example, individuals work in self-managing groups in the absence of a controlling supervisor.[17] Such settings pose monitoring challenges as there is no supervisor available to encourage individuals to act in line with the organizational goals (Druskat & Wolff, 1999; Towry, 2003). Peer evaluations (and by extension peer-based rewards) have been cited as a useful tool for controlling outcomes of self-managing workgroups (Druskat & Wolff, 1999; Huang & Fu, 2013; Saavedra & Kwun, 1993). The use of peer evaluation systems gain in popularity as about 90% of Fortune 1000 firms use some form of peer evaluation nowadays (3D Group, 2013; Carson, 2006; Edwards & Ewen, 1996). Moreover, peer evaluations are even used to (partially) determine personnel decisions regarding promotions and performance pay (Bohl, 1996; Arnold, Hannan, & Tafkov, 2018, 2020). According to the Society for Human Resource Management, 71% of the HR professionals in their sample indicate that the annual performance review process might benefit from including ongoing peer evaluations (SHRM, 2018). Even though peer evaluations are often used in practice, they are heavily contested as peer evaluations might be biased and even disincentive employee effort as peers use them to their own advantage by giving everyone low ratings (Carpenter, Matthews, & Schirm, 2010; SHRM, 2020).

As such scholars have examined the effects of peer performance evaluation systems on employee performance, however little is known about the effects of the design of such systems (Jackson, Michaelides, Dewberry, Schwencke, & Toms, 2020). The way in which firms use a peer evaluation system differs considerably. For example, employees at Google are required to perform peer evaluations twice a year using ratings (Homem de Mello, 2019). These rating systems differ greatly from other firms such as Meta who use forced ranking systems among peers (Gartner, 2018). Also the level of transparency differs. While

[17] Crowdsourcing is a way of collecting knowledge from a bigger group of people (Bayus, 2013) and firms often capitalize on this idea of crowdsourcing to let employees brainstorm in teams to gather ideas for product innovation (Allen et al., 2018; Bayus, 2013) or gather ideas for organization-wide innovation (Gallus et al., 2019; Hodosh et al., 2013; Huang & Fu, 2013; von Ahn & Dabbish, 2004).

Google employees have access to (anonymized) outcomes of peer evaluations (Homem de Mello, 2019), other firms decide to keep such information private (Hannan, McPhee, Newman, & Tafkov, 2013; Tafkov, 2013). Although this variation exists in practice, research to date does not examine the impact of these design features on employee behavior. Besides the practical motivation, it is also important from a theoretical perspective to study why and how features of peer evaluation design affects employee effort. More specifically, the management accounting literature available on the behavioral effects of ratings and rankings (also referred to as forced ratings) remains scarce and focuses on supervisor-employee contexts rather than peer-to-peer contexts (Berger, Harbring, & Sliwka, 2013; Cardinaels & Feichter, 2021). Furthermore, the extent to which outcome transparency affects employee behavior when co-workers have evaluation responsibility has received relatively little research attention to date.

In this study, we examine the effect of transparency for given peer evaluation systems. Transparency in this study relates to the extent that outcomes of peer evaluations are made transparent to peers (Bol, Kramer, & Maas, 2016). We investigate two types of peer evaluations as control systems. Namely, peer evaluations that either require *ratings* or *ranking* from peers. Peer ratings are evaluation systems in which co-workers evaluate each other using a rating scale (e.g. a 9 point Likert scale). In such ratings systems, people are free to give any rating and may put their peers all at low ratings to improve their own position (Carpenter et al., 2010). On the other hand, peer rankings are more restrictive as co-workers rank their peers from best two worst, using a forced distribution rating scale.

We draw on self-concept maintenance theory to develop our predictions (Mazar, Amir, & Ariely, 2008). This theory states that individuals face an ethical dilemma when attempting to promote one's relative standing in a competition. The setting we study is characterized by competition as all participants face a tournament incentive scheme. According to Lazear (1989), individuals can increase their chances of winning the tournament by either exerting high effort or by harming others. The literature on self-concept maintenance conjectures that the extent to which individuals will exert effort or display harming behavior, depends on the extent to which individuals are confronted with their own behavior compared to their peers' (Campbell, Reeder, Sedikides, &

Elliot, 2000; Mazar et al., 2008). Such confrontation can be avoided when peer evaluation outcomes are kept private. Prior work in economics indeed suggests that individuals engage in harming behavior in such contexts (Balafoutas, Czermak, Eulerich, & Fornwagner, 2020; Carpenter et al., 2010; Leibbrandt, Wang, & Foo, 2018). We label this as the *anticipation effect* where individuals expect their peers to underrate them (i.e. engaging in harming behavior) in order to increase their own chances of winning in the competition. This expectation, in turn, disincentives individuals to exert effort.

Drawing on the theory of self-concept maintenance we predict an interaction effect of peer evaluation system and outcome transparency on employee effort. We argue that individuals choose to engage in harming behavior by underrating their peers in peer rating systems, when the outcomes of peer evaluations are not made transparent. The reason is that individuals avoid negatively updating their self-concept in this setting, since they are less likely to be confronted with the costs associated to this behavior. Alternatively, when peer evaluations are based on rankings, we argue that individuals will choose to exert high effort in order to increase their chances of winning the tournament, because harming behavior in this situation will not be beneficial.

However, when peer evaluation outcomes are transparent, peers might experience disutility from appearing as unfair evaluators when evaluations are made transparent to all peers (Maas & Van Rinsum, 2013). Refraining from underrating peers can help them to prevent a negative update in their self-concept. Because of this transparency, we predict that peers will now use ratings to evaluate each other more accurately, which in turn will make the act of exerting high effort more attractive (contrary to engaging in harming behavior). When peer evaluations are based on rankings, we expect that individuals will experience both effortful and harming behavior as costly, because high efforts may not be valued accordingly. Indeed, Berger et al. (2013) find that individuals evaluating others using a forced distribution, express dissatisfaction towards the system as they experience ratings to be difficult. In particular, when peer evaluation information is made transparent, ranking can create more pressure. Scholars suggest that individuals report greater difficulty and lower fairness perceptions under forced distributions rather than free rating scales (Schleicher, Bull, &

Green, 2009). This in turn, can decrease the trust in the system and subsequently may reduce employee effort.

To test our predictions, we conduct an online real effort experiment, using a 2x2 between-subjects design, where participants face a multitask setting in which they have to allocate effort towards quantity and quality. Experimental groups consist of four participants who each take the role of an employee. Participants have to work on an image description task for four rounds. Participants perform the task in a way that their peers can fully observe their output and characteristics that are hard to judge by the principal such as the quality of their descriptions (Carpenter et al., 2010). Employee effort is then operationalized as raw output, measured as the number of images described in each round, for each employee. At the end of each round, participants have to evaluate the quality of their peers' image descriptions. We also measure quality-adjusted output by analyzing whether participants' descriptions match the sourced image captions (Hodosh, Young, & Hockenmaier, 2013), using computer-aided textual analysis. Participants either rate their peers on a 1-9 rating scale in which they are allowed to freely use each rating in the peer rating condition or rate their peers on a one 1-9 scale where they are forced to rank one peer as high performer (7-9), the other as middle (4-6) and the third one as low performer (1-3) in the peer ranking condition (e.g.. Berger et al. 2013; Cardinaels and Feichter, 2021). After participants submit their evaluations, participants either only see their own average received rating (ranking) (not transparent condition), or they also view the average rating (ranking) of all other peers as well (transparent condition). We use average ratings (rankings), an aggregated measure, such that peer anonymity is still warranted.

Our results show that the effect of outcome transparency depends on the type of control system in place. That is, peer rankings can indeed mitigate the anticipation effect connected to peer ratings in settings where the outcomes of the peer evaluations are kept private. More specifically, we find a significant increase in raw output under peer rankings compared to peer ratings, when participants are only informed about their own received average ranking (rating). Additionally, we find a significant interaction effect when analyzing raw output, indicating that the effect of the peer evaluation system depends on the extent to which the outcomes of the peer evaluations are made transparent. More specifically, consistent with our theory, we find that when peer

evaluations are made transparent to all peers, raw output is higher under peer ratings compared to under peer rankings. When we adjust for the quality of participants' image descriptions, we find similar results.

With this study, we want to contribute to the literature in several ways. First, the use of peer evaluations becomes increasingly important, especially in settings where the principals cannot observe how well each individual contributed to team effort, where effort towards a goal is difficult to monitor or where employees work in self-managing groups. Despite the widespread use of such systems in practice, much less is known on how to implement and design these systems so that they can successfully stimulate effort (Jackson et al., 2020; Sol, 2016; Waldman, Atwater, & Antonioni, 1998). We show that when companies decide to keep evaluation information private, the use of peer rankings (instead of peer ratings) increases employee effort. Alternatively, peer ratings (as compared to rankings) increase employee effort when evaluations are made transparent. Second, we contribute to the performance evaluation literature by investigating the effects of peer evaluation systems on employee effort in a competitive setting. While prior research has mainly focused on the effects of different design characteristics of supervisor evaluation (Bol et al., 2016; Cardinaels & Feichter, 2021; Moers, 2005), evidence and theory on the effects of peer evaluations remain scarce.[18] Our study shows that peer rankings can increase effort more than peer ratings when its outcomes are kept private, which is in line with the findings of Berger et al. (2013) who find similar effects of performance evaluation design when supervisors rather than peers rate employees. Additionally, we show that under transparent information policies, participants evaluate their peers more honestly (Evans, Moser, Newman, & Stikeleather, 2016; Maas & Van Rinsum, 2013), which then in turn leads them to exert higher effort. Third, we contribute to Carpenter et al. (2010) by providing evidence that the anticipation effect of peer ratings under non-transparent outcome feedback can be mitigated in two ways. Namely, by installing peer rankings on the one hand, or by making the outcome from the peer ratings transparent to all peers on the other hand. Fourth, given that peer evaluations can be seen as a form of monitoring, we add to the studies examining the effect of peer monitoring in a team context (Falk & Ichino, 2006; Mas & Moretti, 2009; Towry, 2003).

[18] Notable exceptions are Carpenter et al. (2010), and Balafoutas et al. (2020).

Especially in crowdsourcing settings, where multiple individuals are asked for solutions to the same task (Gallus, 2017; Gallus, Jung, & Lakhani, 2019), we show that the effect of peer monitoring depends on the system in place. Specifically, peer ratings as a form of monitoring can enhance the quality of crowdsourced outcomes when information is made transparent but not when evaluation outcomes remain opaque.

2.2 Background and hypothesis development

2.2.1 Background

In this study, we examine the effects of peer evaluation system and outcome transparency in a competitive setting. That is, in our multi-period real-effort experiment, participants work on a task in which they have to allocate effort both to quantity and quality. In all conditions, participants can earn a bonus if they outperform their peers, on top of their piece-rate compensation. These types of tournament incentives induce competition between participants in order for them to win the tournament (i.e. the bonus) (Lazear & Rosen, 1981).

2.2.1.1 Peer evaluations in prior literature

A first strand in the literature on peer evaluations focuses on the positive effects of peer evaluation from an agency theoretic point of view. Agency theorists show that peer evaluations are an incentive for individual effort and performance (Marx & Squintani, 2009; Sol, 2016). The authors argue that when principals are unable to observe individual efforts in a team work environment, the mere fact of peers evaluating each other incentivizes individuals to exert effort. The reason is that the probability of detection for shirking or free-riding is greater when peers monitor each other, and agent shirking can therefore be penalized (Marx & Squintani, 2009). Research has shown that increased communication and cooperation between members and less free-riding can explain the positive relationship between peer evaluation and performance (Erez, Lepine, & Elms, 2002; Towry, 2003). Agency theorists even argue that these effects extend to settings where individuals do not report truthfully about their peers (Sol, 2016).

A second body of the literature focuses on the effects of the mere presence of peers. Scholars have argued that peer effects exist based on

spatial proximity (Falk & Ichino, 2006; Mas & Moretti, 2009). Falk and Ichino (2006) show that individuals perform better in a letter stuffing task when they are put in a room together with a peer. Similarly, Mas and Moretti (2009) find that worker productivity in a supermarket chain is higher when workers observe each other. Chen and Sandino (2012) extend these results in their field study, where they find that employee theft in retail chains is lower when wages are high because they promote social norms among co-workers. These results can be explained by the premise that individuals prefer to be approved by their peers and have an adversity for disapproval (López-Pérez & Vorsatz, 2010).

While the previous studies model peer evaluation as either 'evaluation messages' (Sol, 2016) and 'reports' (Marx & Squintani, 2009) or mere presence of peers (Chen & Sandino, 2012; Mas & Moretti, 2009), little is known about how the design of these evaluations affects employee effort. We, therefore, investigate a peer setting in which individuals can observe each other's work and consequently perform an evaluation on their peers using either peer ratings or peer rankings as different systems of peer evaluations. Peer ratings consist of individuals being rated by their group members on a given set of performance (and/or personality) characteristics (Kane & Lawler, 1978). Peer rankings on the other hand, generally require individuals to rank their peers from best to worst on a given performance dimension similar to forced distribution systems (Berger et al., 2013; Cardinaels & Feichter, 2021).

2.2.1.2 The anticipation effect of peer ratings

Although agency theory thus predicts an incentive effect of peer evaluations, this is in sharp contrast to what behavioral economists find (Balafoutas et al., 2020; Carpenter et al., 2010; Leibbrandt et al., 2018). That is, ratings often suffer from rater biases. There is evidence that peer ratings can be biased (DeNisi et al., 1983; Fedor et al., 1999; Ahn, Hwang, & Kim, 2010; Bol, 2011; Rosaz & Villeval, 2012). More specifically, peer ratings can suffer from self-enhancement bias, which involves errors in (peer) evaluations stemming from the personal motivation to maintain and increase one's self-image (Alicke & Sedikides, 2009). Behavioral economists further suggest that individuals expect these ratings to be biased and this in turn can have an effect on individual effort.

In an experiment by Carpenter et al. (2010), the authors find that when individuals are being evaluated by means of peer ratings, they expect the ratings to be biased. More specifically, peers may underrate each other's performance in order to increase their own chances of winning a bonus (which could be seen as a form of self-enhancement bias, see Alicke and Sedikides, 2009). Carpenter et al. (2010) argue that this expectation causes the marginal benefit of effort exertion to shift down, resulting in decreases in effort and output. Likewise, Balafoutas et al. (2020) find the same effects of peer ratings in an experiment with junior auditors. We refer to this effect as the *anticipation effect* of peer ratings, where individuals anticipate biased reflections of their performance from their peers, which in turn discourages them to exert effort. It is important to note that evidence for this anticipation effect is only found when peer ratings remain strictly private.[19]

Therefore, this study attempts to replicate and extend the (inconclusive) findings of the peer evaluations established in prior research by considering the role of outcome transparency for given peer evaluation systems. The focus on transparency of the peer evaluation stems from a growing body of research suggesting that control mechanisms where one can view *outcomes* of (peer) evaluations can have a significant impact on employee behavior (Abeler, Falk, Goette, & Huffman, 2011; Bol et al., 2016; Falk & Ichino, 2006; Hannan, Towry, & Zhang, 2013). We follow Bol et al. (2016, p. 66) who define outcome transparency as "the extent to which employees have access to information on the outcomes of the evaluation processes…". Evidence indeed shows that transparency on evaluation policies can vary significantly across organizations (Colella, Paetzold, Zardkoohi, & Wesson, 2007; Futrell & Jenkins, 1978; Hannan et al., 2013; Lawler, 1990; Tafkov, 2013), where some companies decide to only display the final evaluation outcome, and others make the evaluation process more transparent.

[19] Balafoutas et al. (2020) even argue that underreporting peers' output can be classified as unethical behavior, referring to it as a form of sabotage (Lazear, 1989). Indeed, the fear for peers behaving unethically is grounded since Tzini and Jain (2018) find that under relative performance evaluation, individuals expect others to be more likely to behave unethically and they indulge more into unethical behavior themselves.

2.2.1.3 Theory of self-concept maintenance and outcome transparency

Individuals are primarily motivated to avoid decrements in their concept of self, which can be achieved by promoting one's relative position (Baumeister, Bratslavsky, Finkenauer, & Vohs, 2001). To promote one's relative position, individuals face an ethical dilemma between gaining financial benefits from harming others and maintaining a positive self-concept (Aronson, 1969; Harris, Mussen, & Rutherford, 1976). The theory of self-concept maintenance then postulates that individuals are more likely to pursue these financial benefits from behaving unethically (i.e. harming others) when they can somehow still maintain a positive self-concept (Mazar et al., 2008). Indeed, individuals strive to enhance their selves and are less afraid to give in to this self-serving bias (Alicke & Sedikides, 2009), when outcomes are not made transparent. That is, individuals may underrate peers to look good in the competition.

When individuals are aware of their peers' outcomes, rewards and ratings, their behavior can be altered. The theory of self-concept maintenance suggests that comparison between oneself and peers can influence the extent to which individuals engage in harming behavior (i.e. self-serving bias). That is, employees in transparent settings will improve one's relative position by exerting effort as they want to maintain a positive self-concept. Furthermore, prior literature has shown that individuals in transparent settings are less likely to display self-serving bias as honesty and fairness are social norms[20] that most individuals avoid breaking publicly (Bicchieri, 2005; Fehr & Gachter, 2000). That is, when peer evaluations become transparent, individuals might experience disutility from breaking social norms publicly.[21] This disutility then materializes in negatively updating one's self-concept and

[20] Social norms are defined as 1) a behavioral regularities; that are 2) based on a socially shared beliefs of how one ought to behave; which trigger 3) the enforcement of the prescribed behavior by informal social sanctions (Fehr & Gachter, 2000 p. 166).

[21] In a study by Maas & Van Rinsum (2013), the authors argue and show that individuals experience disutility when they are being perceived as dishonest by their peers. Their findings indicate that performance reporting under a transparent information policy is more honest as compared to a non-transparent information policy (Maas & Van Rinsum, 2013). Similarly, Cardinaels & Jia (2015) show that honesty in reporting increases with transparency if combined with audits. Finally, consistent with social norm theory, Chen & Sandino (2012) show that employee theft decreases when co-workers are present in a retail chain and when wages are relatively high.

experiencing disapproval of one's behavior by peers (López-Pérez & Vorsatz, 2010).

2.2.2 Hypothesis development

As explained above, our study setting is characterized by competitiveness. Prior literature on tournament incentives has shown that tournaments increase both productive and counterproductive efforts (Berger et al., 2013; Charness et al., 2014; Harbring & Irlenbusch, 2011; Lazear (1989)). Individuals have two strategies to win in a tournament. Namely, (1) they can increase their own effort, or (2) they can win by harming others. To develop our hypothesis we rely on the theory of self-concept maintenance to predict which of the above two strategies dominates depending on the extent to which the outcomes of two distinct peer evaluation systems are made transparent.

2.2.2.1 The role of peer evaluation systems

Employees performing peer evaluations essentially have two roles, namely the role of evaluator and the role of being the evaluated one. As prior research on peer evaluations describes, each role comes with costs and benefits (Fedor, Bettenhausen, & Davis, 1999), and it is shown that peers consciously consider their rating behavior in weighing the costs and benefits w.r.t. peer evaluations (Ng, Koh, Ang, Kennedy, & Chan, 2011; Spence & Keeping, 2010). That is, when individuals carry out peer evaluations, they deliberately think about the consequences of their role as an evaluator and how this will impact the evaluatee. This, in turn, leads individuals to expect that their peers will behave in the same way (Mazar et al., 2008). Therefore, we expect individuals to take into account each other's evaluation behavior, by reasoning backward.

According to the self-concept maintenance theory, employees' self-concept is less likely to be updated when they are not confronted with their own behavior. Mazar et al. (2008) argue that this is the case when individuals are *inattentive* to social norms, which inhibits them to evaluate their actions in light of the norm. Based on this logic, individuals are expected to choose the strategy of harming others (underrating peers in our setting) in order to increase their chances of winning the tournament bonus, when peer evaluation outcomes are not transparent as in this case individuals are not confronted with their own behavior and

that of their peers. The extent to which the strategy of harming others is beneficial depends on the possibilities to do so (i.e. the peer evaluation system in place).

In the peer rating condition, we predict that individuals will anticipate that peers underrate each other in order to increase their own payoffs, because of their dual role as evaluator and evaluatee. In other words, we expect to replicate the *anticipation effect* of Carpenter et al. (2010). Consequently, individuals are discouraged to exert effort in the first place since they expect their received ratings to be biased (Ahn et al., 2010; Balafoutas et al., 2020; Berger et al., 2013; Bol et al., 2016; Carpenter et al., 2010; Leibbrandt et al., 2018; Taggar & Brown, 2006; Rosaz & Villeval, 2012).

In the peer ranking condition, we expect that individuals engage in the harming strategy to a lesser extent, because rankings force evaluators to differentiate their ratings (Berger et al., 2013; Cardinaels & Feichter, 2021). While it is not clear upfront how the effect of peer rankings would manifest in a peer evaluation setting, we still expect increased employee effort in this condition (relative to peer ratings). Under a ranking system, evaluators have to rank their peers from best to worse (Berger et al., 2013; Kane & Lawler, 1978) and they cannot give each group member the same rating. Because of these restrictions, they are less able to subsequently underrate other peers' performance. Consequently, individuals will anticipate that peers - because of the nature of evaluation - will not engage in underrating, which implies that ratings might be more accurate. In this condition then, it is most beneficial to increase one's own effort in order to increase one's chances to win the tournament bonus.

Consistent with this theory, prior research finds that the provision of relative performance ranks (Hannan, Krishnan, & Newman, 2008; Hannan, McPhee, et al., 2013; Tafkov, 2013) and the use of forced distribution ratings rather than free ratings (Berger et al., 2013) increase employee effort. In a similar vein, Cardinaels and Feichter (2021) show that this incentive effect applies to settings where performance is evaluated objectively (rather than subjectively) by supervisors. Building upon this evidence, we propose the following (partial) hypothesis:

Hypothesis 1a: *When peer evaluation outcomes are not transparent, effort levels in peer rating systems are lower than effort levels in peer ranking systems.*

However, when outcomes of peer evaluations are publicly available, individuals are more likely expected to comply with social norms. We expect that under a transparent setting, individuals will more often choose the strategy of effort exertion rather than harming others when peer evaluations are done through peer ratings. This could in turn serve as an effort incentive, where individuals expect others to behave in the same way (because of their dual role as evaluator and evaluatee). That is, underrating each other implies breaking the norm of giving a fair evaluation, and hence individuals might be less incentivized to do so when average received ratings are public, because this will lead them to negatively update their self-concept (Mazar et al., 2008). Indeed, in this condition individuals care about appearing honest and fair with their peers, and this transparency thus provides an incentive for honest and fair peer evaluations (López-Pérez & Vorsatz, 2010). If individuals engage in harming behavior in this setting, there is also the cost of possible retaliation. Therefore, individuals will reason that it is more beneficial to engage in effort exertion behavior compared to harming behavior.

When peer evaluations use rankings in transparent settings, individuals may reason that effort exertion is again a dominating strategy. However, in forced distribution or ranking systems, there is a risk that high effort exertion is not valued by the evaluator. The transparency of the rankings can help individuals to assess how they rank against others and whether such ranking accurately captures their performance. Indeed, prior evidence in transparent settings shows that individuals perceive ranks as less fair than ratings (Roch, Sternburgh, & Caputo, 2007). Evaluating peers using forced ranking buckets can be experienced as difficult (Berger et al., 2013; Scheichel et al., 2009) and when evaluations are made transparent such difficulty in ranking becomes even more prominent. These concerns then disincentive individuals to exert high effort. Hence our second (partial) hypothesis is as follows:

Hypothesis 1b: *When peer evaluation outcomes are transparent, effort levels in peer rating systems are higher than effort levels in peer ranking systems.*

Together H1a and H1b predict a disordinal interaction between peer evaluation system and outcome transparency. More specifically, we predict that the dominating strategy for increasing one's chances of

winning the tournament bonus depends on both the peer evaluation system and the extent to which outcomes are made transparent.

Hypothesis 1: *Outcome transparency and peer evaluation system interact such that under a transparent setting, individuals in the peer rating condition exert higher effort than individuals in the peer ranking condition.*

2.3 Experimental method and design

We test our hypotheses in a 2×2 between-subjects experiment in which we manipulate the system used for evaluating one's peers (i.e. a *peer ranking* versus *a peer rating* system), and the extent to which participants receive feedback on these peer evaluations (i.e. *transparent* versus *not transparent*). The experiment is coded using oTree, a Python-based framework for conducting online interactive experiments (Chen, Schonger, & Wickens, 2016). In total, 364 participants from two large European universities took part in a compensated online interactive experiment.[22,23] Participants received an average payment of €9 for approximately 48 minutes of participation time. Male students represented 46.39% of the sample. Participants in the study were on average 21.56 years old and they had about 17.39 months of work experience.

2.3.1 Experimental task and procedure

Similar as in previous studies that examine peer evaluation, we use a real-effort task that does not require specific skill sets (Ariely, Kamenica, & Prelec, 2008; Berger et al., 2013; Carpenter et al., 2010, Falk, & Ichino, 2006). The benefit of using a real effort task is that performance is

[22] Data collection proceeded in two stages. In the first round, 160 students were recruited from a large Belgian university in December 2020. In the second round, another 204 students from a large Dutch university were recruited in February 2021. Controlling for the sample difference does not change the results for our hypothesis. The study received ethical approval from both universities.

[23] After data cleaning, we deleted 32 observations due to technical issues (16 observations from the Belgian sample, and another 16 from the Dutch sample), resulting in a total of 332 observations left for data analysis.

reflected by the amount of effort individuals put in the task and less by ability and skill. Additionally, our experimental task resembles a crowdsourcing task where peers perform image descriptions and need to evaluate the quality of the descriptions. As image descriptions are used as input for machine learning (teaching computers to translate images into natural language) (Hodosh et al., 2013; Huang & Fu, 2013; von Ahn & Dabbish, 2004), they often require input from multiple employees who describe the same image. The task is similar to the letter-stuffing task used by Carpenter et al. (2010) in a way that we can distinguish between quantity and quality, where quantity is precisely observable, while quality not. Moreover, it is a task where peer evaluations are useful, since peers develop the experience of writing image descriptions while doing the task. This way, peers are able to judge the quality of their peers (which might be difficult without experience) (Kane & Lawler, 1978).

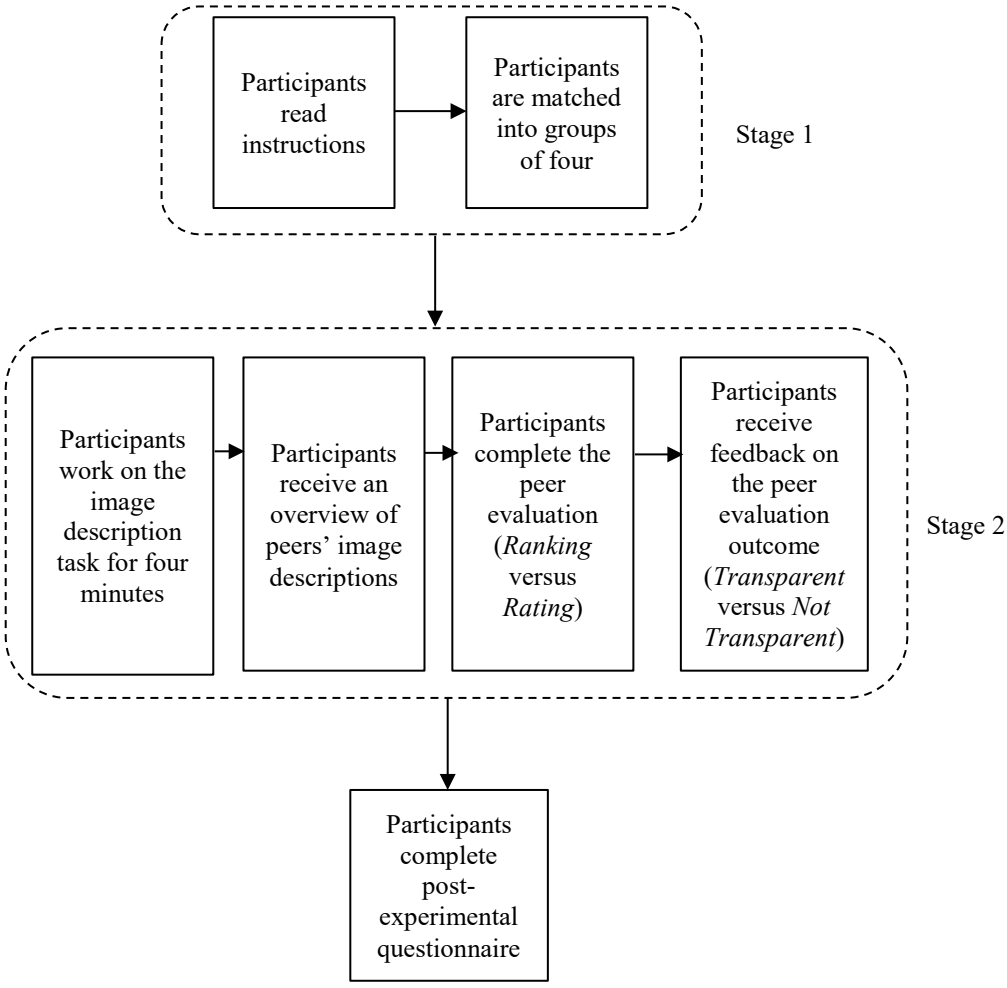
In all conditions, participants assume the role of an employee whose job is to produce image descriptions and they are informed on the rules of the peer evaluations and how their payoff is determined based on these evaluations. Figure 2.1 offers an overview of the experimental timeline. To ensure that every participant understands the instructions, participants take a quiz on how the peer evaluation system works and how their payoffs are determined. After finishing the quiz, participants are sent to wait in a virtual waiting room. Once four participants are present in the waiting room they are assigned to one group, after which they start to describe and evaluate images in the same cohort of four (stage 1).[24]

Participants work for four work periods lasting 4 minutes, in which they have to describe a different set of images. Each group member is presented with the same image sequences. At the end of each work period, an overview screen displays the image descriptions of the participants' peers. In the subsequent screen, participants complete and submit the peer evaluation. The peer evaluation is set up such that each participant is evaluated based on the quality (i.e. detail and accuracy) of their image descriptions. Right after submitting the peer evaluation, a feedback screen shows the outcome of the peer evaluation (i.e. the received average ratings/rankings). The content of the feedback screen depends on the condition. This sequence of activities constitutes stage 2 of the study and is thus repeated four times. At the end of the study,

[24] To avoid lengthy waiting time, group matching is based on participants' arrival time after they finish reading the instructions and quiz (on a first come, first served basis).

participants completed a post-experimental questionnaire that allowed us to collect their thoughts during the experiment, as well as the usual demographic information.

Figure 2.1: Overview of the experimental timeline



2.3.2 Independent variables

2.3.2.1 Peer evaluation system and payoff

The first independent variable is the peer evaluation system, manipulated

either as a rating system or as a ranking system, also referred to as a forced ranking (Berger et al., 2013; Cardinaels and Feichter, 2021). Employees are asked to evaluate the *output quality* for each of their peer group members on a scale from 1-9 (1=Below Average, 9=Above Average with 5 as the midpoint). Participants have no information about the identity of their peers.[25] We opt for this design choice in order to minimize conditional reciprocity in rating behavior (Balafoutas et al., 2020) because if participants might know each other, this could influence their rating behavior (i.e. giving favorable ratings to individuals they know). In the peer *rating* condition employees are not restricted in their rating behavior. In the peer *ranking* condition, however, employees are explicitly instructed that they have to rank each of their peers differently. More specifically, employees have to give one peer a rank of 1-3 (Below Average), another peer a rank of 4-6 (Average), and another peer a rank of 7-9 (Above Average). The restriction in the *ranking* condition is further ensured to the employees by pop-up error messages whenever a participant fails to differentiate his/her rankings. The 1-9 range allows us to ensure comparability across the *rating* and *ranking* conditions, without losing granularity for allowing differentiated ratings.

The outcome of the peer evaluations determines the participant's payoff. Participants are informed that their payoff is not only determined by the number of image descriptions produced (i.e. output) but also by the quality of their descriptions. Each employee's payoff in each work period takes the following form:

$$Pay = N \times Q \times 100\ lira$$

where N is the count of the production output (in a specific work period), Q is the *quality-adjustment parameter*, and lira is the experimental currency. For each quality-adjusted unit (N × Q) produced, participants earn 100 lira.[26]

The calculation of Q (quality-adjustment parameter) is based on a participant's *average received rating/ranking*. We calculate participant's average ratings/rankings as the sum of all ratings divided by three (i.e. the number of peers evaluating a given participant's output quality).

[25] The participants in our study are asked to create their own unique nicknames. These nicknames are then displayed to fellow group members when participants see the outcome, peer evaluation, and feedback screen.

[26] The exchange rate from lira to Euro is determined such that all (total) payoffs fall in a range between €6 and €18.

When a participant's average received rating/ranking is higher than 6.5 the Q for calculating his/her payoff is 1, for a rating/ranking between 3.5 and 6.5 Q equals 0.5 and when the average rating/ranking is below 3, Q equals 0.25.

In addition to the piece rate payoff determined by the above formula, employees can earn a tournament bonus of 100 lira in each round. This way, the compensation scheme is competitive in nature. The bonus is awarded to the employee with the highest quality-adjusted output (N × Q) in his/her group.

2.3.2.2 Outcome transparency

The second independent variable we manipulate is whether or not next to the individual's average rating, the average peer evaluation outcomes are transparent or opaque. In the transparent condition, employees are displayed an overview of the average ratings/rankings from all group members, while in the *non-transparent* condition employees only see their own average received rating/ranking. Our manipulation follows Bol, Kramer & Maas (2016) where we inform the participants that the peer evaluations are made publicly available to all employees (kept strictly confidential) such that all group members see each other's average ratings/rankings (each employee only sees his/her own average rating), respectively for the *transparent* (*non-transparent*) condition.[27]

2.3.3 Dependent variables

The main variable of interest is *effort* in this study. Since the experimental task in this study is a real-effort task where participants work for fixed time periods, we proxy effort by the *raw output* (N) of image descriptions produced (Carpenter et al., 2010, Falk, & Ichino, 2006). Admittedly, individuals work in a multitask setting where they have to allocate effort towards both quantity and quality. As prior research shows, the extent to

[27] The participants were informed that these average received ratings/rankings were aggregated, such that participants did not see the decomposed ratings/rankings received from each group member. This was again done in order to minimise conditional reciprocity. This manipulation is further strengthened by reminding participants that the results of the peer evaluation will be made public (kept confidential) in the transparent (non-transparent) condition, right after each work period.

which individuals focus on the quantity or quality side of output depends on how both are incentivized (Rubin, Samek, & Sheremeta, 2018). Similar to Carpenter et al. (2010), individuals in this study are compensated based on their quality-adjusted output, taking into account both quantity and quality. As such, we expect individuals not to engage in a strategy of producing a high number of low-quality image descriptions, as doing so imposes risks for their compensation. Hence, our measure of raw output constitutes a conservative proxy for effort.

While we have not proposed hypotheses on effort allocations towards output quantity or quality, we do analyze *quality-adjusted output* $_{Objective}$ as an additional dependent variable. We measure participants' quality-adjusted output by performing computer-aided textual analysis on their image descriptions based on the image captions from the Flickr-8K database (Hodosh et al., 2013). This database contains over 8.000 images along with five crowdsourced image captions. From these captions, we first compiled a list of keywords for each image. In a next step, we identified the number of keywords in each participant's i image description j, based on our keyword list. To construct a measure reflecting the quality of each image description, we calculated the percentage of identified keywords against the total number of unique keywords in the Flickr-8K caption for each participant's i image description j.[28] This percentage then, constituted our externally validated quality-adjustment parameter Q. We calculate *quality-adjusted output* $_{Objective}$ (N×Q), using the Q below.

$$Quality-adjustment\ \ parameter_{i,j} = \frac{N\ of\ keywords\ identified\ in\ a\ participant's_i\ image\ description_j}{N\ of\ unique\ keywords\ in\ Flickr-8K\ captions\ of\ image_j}$$

Similar to Carpenter et al. (2010) who report results on quality-adjusted output by simply multiplying the raw output with the respective quality-adjustment parameter from the experiment, we report *quality-adjusted output* $_{Subjective}$ as (N×Q), using the Q resulting from the peer evaluations from the experimental study. We label this measure as subjective, since it is constructed based on the (subjective) *ratings* and *rankings* from the

[28] By *unique* keywords we represent one word that can be conjugated differently (i.e. *skateboarder* and *skater* are considered as one unique keyword, and *watching* and *watches* are also considered as one unique keyword).

peer evaluations in our experiment, which can be subject to rater bias as described earlier.

2.4 Results

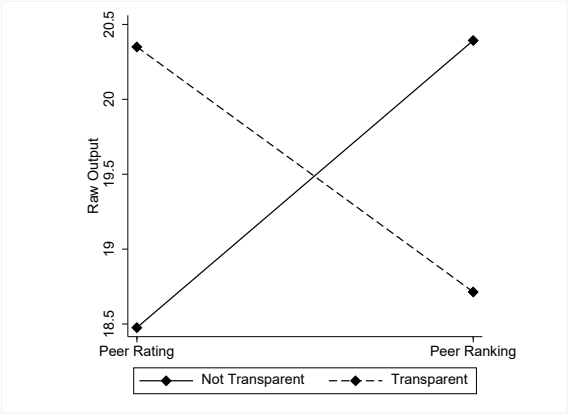
2.4.1 Descriptive statistics, manipulation checks and randomization check

Descriptive statistics for the dependent variables to test our hypotheses are presented in Table 2.1, and displayed graphically in Figure 2.1. The correlations between all variables are presented in Table 2.2.

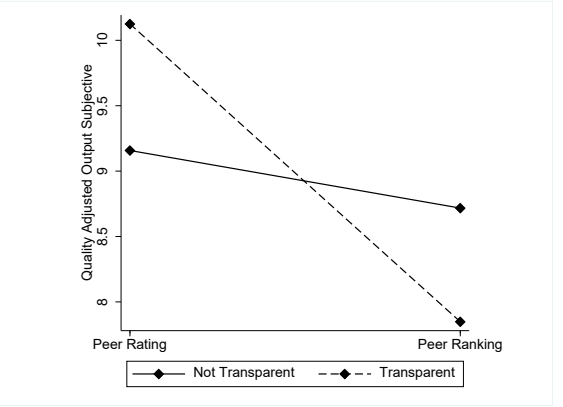
Figure 2.2: Graphical representation of dependent variables

Graphical depictions for the raw output over all work periods (Panel A), and the quality-adjusted output $_{\text{Subjective}}$ over all work periods (Panel B), by condition.

Panel A:



Panel B:



Panel A shows the mean number of image descriptions produced (N) by participants, by condition.
Panel B shows the mean number of image descriptions produced by participants multiplied by their *quality-adjustment parameter* (N × Q) (which results from the peer evaluation), by condition.
Note that the raw output, and quality-adjusted output variables are aggregated over all four work periods.

Table 2.1: Descriptive statistics

		Peer Ranking		Peer Rating	
		Not Transparent	**Transparent**	**Not Transparent**	**Transparent**
Panel A: Individual-level observations					
Raw individual Output	Work period 1	3.464 (2.432)	2.988 (1.697)	3.250 (2.233)	3.662 (2.365)
	Work period 2	5.000 (2.364)	4.333 (2.251)	4.405 (2.689)	5.025 (2.455)
	Work period 3	5.726 (2.495)	5.464 (2.346)	5.297 (2.692)	5.662 (2.495)
	Work period 4	6.202 (2.692)	5.929 (2.358)	5.524 (2.682)	6.000 (2.413)
	Total over all periods	20.393 (8.750)	18.714 (7.227)	18.476 (9.366)	20.350 (8.836)
Average received rating	Work period 1	5.151 (1.247)	5.151 (1.588)	5.825 (1.213)	5.788 (1.352)
	Work period 2	4.976 (1.497)	5.028 (1.724)	5.611 (1.355)	5.654 (1.492)
	Work period 3	4.940 (1.485)	4.952 (1.807)	5.560 (1.428)	5.588 (1.436)
	Work period 4	4.956 (1.527)	4.893 (1.593)	5.341 (1.567)	5.5208 (1.372)
Quality-Adjusted Output $_{Subjective}$		8.717 (3.917)	7.848 (3.391)	9.158 (5.584)	10.125 (5.577)
Quality-Adjusted Output $_{Objective}$		3.308 (1.249)	3.161 (1.186)	3.098 (1.198)	3.486 (1.278)
Number of observations		84	84	84	80
Panel B: Group-level observations					
Raw Group Output		81.571 (26.374)	74.857 (17.962)	73.905 (29.271)	81.400 (26.425)
Number of groups		21	21	21	20

This table presents the mean, (standard deviation) for the dependent variables over all work periods, for each condition.

Table 2.2: Pearson correlations

	1	2	3	4
1 Raw Output	1.000			
2 Raw Group Output	0.742***	1.000		
3 Quality-Adjusted Output Subjective	0.778***	0.618***	1.00	
4 Quality-Adjusted Output Objective	0.383***	0.263***	0.301***	1.00

*,**,*** denote significance at the 10%, 5%, and 1% level, respectively.

To assess the effectiveness of our first manipulation *peer evaluation system*, participants had to indicate the extent to which they agreed with the following statements measured on two 1-7 Likert scale items (1 = strongly disagree, and 7 = strongly agree, with 4 as the midpoint): (1)"When I had to assess my peers, I did this by rating each of my peers on a scale form 1-9 based on their output quality, without being required to distinguish my ratings" and (2) "When I had to assess my peers, I did this by giving them each a different ranking, one in the 1-3 range, another in the 4-6 range and the other in the 7-9 range".

The mean response for item (1) in the *ranking* condition (M=4.42, SD=2.33) was lower than the mean response for the *rating* condition (M=5.47, SD=1.37). The difference between both means is significantly different from zero (t_{330}=-4.973, p-value<0.001 two-tailed). The mean response for item (2) in the *ranking* condition (M=6.72, SD=0.64) was higher than for the *rating* condition (M=2.57, SD=1.17). The difference between both means is again different form zero (t_{330}=30.31, p-value=0.000 two-tailed). These results suggest that our *peer evaluation system* manipulation was successful.

To assess the effectiveness of *outcome transparency* manipulation we asked the following items: (3)"When I submitted my peer assessment, I thought that the other employees in my group would get to know my reported evaluations" and (4) "My submitted peer assessment was unknown to the other employees in my group". The mean response for item (3) in the *non-transparent* condition (M=3.03, SD=1.85) was lower than the mean response for *transparent* condition (M=4.05, SD=1.88), and the difference between the two means is significantly different from zero (t_{330}=-4.99, p-value<0.001 two-tailed). Additionally, the mean

response for item (4) in the *non-transparent* condition (M=5.56, SD=1.62) was higher than the mean response in the *non-transparent* condition (M=4.68, SD=1.79) and this difference between the two means was again significant (t_{330}=4.72, p-value<0.001 two-tailed). These combined results indicate that the manipulation for feedback transparency was successful.[29]

In order to check random assignment of participants to the conditions we conduct a series of analyses. We do not find significant effects of our manipulations for the variables gender, sample (i.e. Dutch versus Belgian students) and prior work experience (all two-tailed p-values>0.21). However, when considering age as dependent variable, we find a significant effect of the *peer evaluation system* ($F_{1,331}$=3.33, p-value=0.07 two-tailed). Participants in the peer ranking conditions (M=21.98, SD=4.70) were slightly older than participants in the peer rating conditions (M=21.13, SD=3.62). Likewise, when considering study levels as the dependent variable, we find a significant effect of the *peer evaluation system* ($F_{1, 330}$=5.77, p-value= 0.02 two-tailed). Further results suggest that 32.3% of the participants in the peer ranking condition were enrolled in master programs, whereas only 19.5% of the participants in the peer rating condition were enrolled in master programs. All remaining students are mainly bachelor students. Controlling for both worker age and worker study level, does not change the results for our hypotheses.[30]

2.4.1.1 Rating behavior

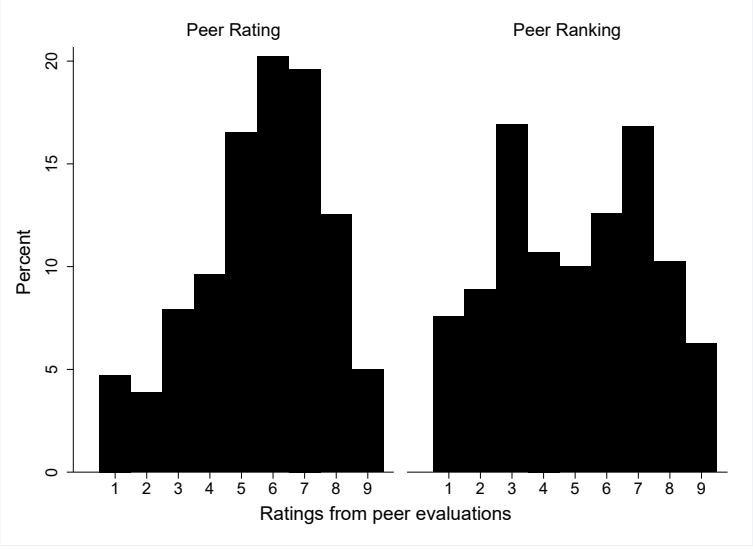
In addition to our manipulation tests, we analyze participants' rating behavior for each peer evaluation system. For each of the two peer evaluation conditions *rating*, and *ranking* we thus have 41 and 42 independent group observations, respectively. Figure 2.3 depicts the

[29] It must be noted that the mean responses on both statements are quite similar (i.e. especially for the fourth statement, both mean responses are around the midpoint of the 7-point Likert scale). The reason is that participants might have experienced transparency as less clear-cut compared to the peer evaluation system they were using. Nevertheless, recall that right after each work period, participants were reminded of their respective treatments w.r.t. the outcome transparency, which strengthens our manipulation even further.

[30] If we alternatively code study level as a binary variable indicating 1 when participants are enrolled in a master program and 0 otherwise, results for our hypothesis also remain the same.

distribution of ratings over all four working periods for the *rating* and *ranking* conditions. Participants in the *rating* condition (see Figure 2.3, left panel), tend to assign ratings that are average or above average, i.e. rating in the range of 5 to 8, in the majority of cases (64%). That is, the distribution of ratings in the *rating* condition is more centered to the right. Alternatively, Figure 2.3 (right panel) shows that the ratings in the *ranking* condition are more dispersed over the total 1-9 rating scale, suggesting that participants in the *ranking* conditions assign more discriminatory ratings than participants in the *rating* conditions. In addition, a chi-square test revealed that the distributions of ratings for the *rating* and *ranking* conditions differed significantly ($X^2_{8,3985}$ = 196.91, p-value<0.001 two-tailed).

Figure 2.3: Rating behavior across peer evaluation system



2.4.2 Hypothesis tests

2.4.2.1 Hypothesis 1a

Hypothesis 1a predicts that when the outcomes of the peer evaluations are not transparent to employees, effort levels in peer ranking systems are higher than effort levels in peer rating systems. Note that we proxy effort levels by *raw output.* Table 2.1 presents the means for raw output

for the conditions of interest. Visual inspection of these means shows that the mean raw output in the *ranking* condition (M=20.39, SD=8.75) is higher than the mean raw output in the *rating* condition (M=18.48, SD=9.37), suggesting preliminary evidence for H1a. Indeed, we find further marginal support for H1a by formally testing whether the difference between the latter means is greater than zero (t_{166}=1.37, one-tailed p-value=0.09). These results are consistent with H1a, suggesting that the anticipation effect found in prior literature on peer evaluations, can be mitigated by having peers evaluate each other in a ranking system, instead of a rating system, only when the outcomes of such peer evaluations remain private. Note that the differences in effort levels are manifested already during the first working period (see Table 2.1 for a breakdown of raw output over all four working periods).

2.4.2.2 Hypothesis 1b

Hypothesis 1b predicts that when the outcomes of the peer evaluations are made transparent to employees, effort levels in the peer rating condition are higher than effort levels in the peer ranking condition. We refer again to Table 2.1, where the means of raw output are presented. As can be seen from Table 2.1, the mean raw output in the *Rating* condition (M=20.35, SD=8.84) is higher than the mean raw output in the *Ranking* condition (M=18.74, SD=7.23) when the outcomes of the peer evaluations are made transparent to all peers. Formally testing the difference between the latter two means, suggests that H1b is also marginally supported (t_{162}=-1.30, one-tailed p-value<0.10). Hence, we find evidence that when peer evaluations are made transparent, individuals exert more effort when they evaluate their peers using a rating system rather than a ranking system.

2.4.2.3 Hypothesis 1

Recall that H1a and H1b together predict the interaction effect between the peer evaluation system and the feedback transparency, such that the effect of the peer evaluation system on employee effort will depend on the extent to which feedback on the outcomes of the peer evaluation are made transparent or not. That is, we expect that when peer evaluation outcomes are not made transparent, peer ranking systems increase effort more than peer rating systems (see H1a), while we expect this effect to be reversed when the peer evaluation outcomes are made transparent (see

H1b). To test H1, we conduct a series of regressions on raw output, see Table 2.3.[31] In the first specification we compute the average individual output over all four working periods and regress it on our treatment dummies. In the second specification then, we use a random effects regression model to regress all individual observations over all working periods (i.e. raw output for each individual during each working period) controlling for the time trend by including period dummies. In the third specification we use group observations in all working periods (i.e. the sum of all group members' raw output) as the dependent variable. Table 2.3 demonstrates the results of these regressions (as in Berger et al., 2013).

Column (1) shows a significant interaction effect of the peer evaluation system and feedback transparency. That is, when individuals are being evaluated by means of a peer rating system, their raw output increases when the peer evaluations are made public than when they remain private. The coefficients obtained in model (2) shows that this result remains unchanged. These results thus, provide evidence consistent with H1, showing that a transparency effect exists such that participants exert higher effort when they are being rated by their peers. The simple effects for model (2) show that the effect of peer evaluation is significant in non-transparent feedback settings (β=-0.88; p-value=0.06 two-tailed, untabulated). Column (3) then shows the estimated coefficients when clustering observations on *group_id*. Although the sign of the interaction term is negative, it is not significant.[32]

[31] We base our analyses on the approach taken by Berger et al. (2013), and Cardinaels & Feichter (2021).

[32] The one-tailed p-value however, is close to the conventional cut-off levels of significance. The coefficient on the interaction term of model 3 suggests that individuals under a peer rating system produce less image descriptions when outcomes are not made transparent compared to when they are made publicly transparent (one-tailed p-value=0.105).

Table 2.3: Test of hypothesis

Dependent variable:	***Raw Output***			***Quality − Adjusted Output*** $_{Objective}$		
	(1) OLS	(2) RE (individuals)	(3) RE (groups)	(4) OLS	(5) RE (individuals)	(6) RE (groups)
$Transparent$	-1.679	-0.420	-0.420	-0.147	-0.037	-0.037
	(1.324)	(0.331)	(0.498)	(0.189)	(0.047)	(0.051)
$Rating$	1.636	0.408	0.410	0.325*	0.081*	0.081
	(1.340)	(0.471)	(0.504)	(0.192)	(0.048)	(0.052)
$Not\ Transparent$	-3.552*	-0.888*	-0.888	-0.535**	-0.134**	-0.134*
$\times\ Rating$	(1.884)	(0.471)	(0.709)	(0.270)	(0.067)	(0.073)
Constant	20.393***	6.143***	6.143***	3.308***	0.926***	0.926***
	(0.936)	(0.242)	(0.364)	(0.134)	(0.035)	(0.040)
Observations	332	1328	1328	332	1328	1328
Number of groups/subjects	332	332	83	332	332	83
Adj. R^2 / Wald X^2	0.002	754.57	363.48	0.006	570.41	280.89

Robust standard errors in parentheses *,**,*** indicate p-values at the 10%, 5%, and 1% level respectively.
Robust standard errors are in parentheses (and in (3) and (6) clustered on *group_id*). In columns (2), (3), (5), and (6), period dummies are included. Columns (1) and (4) show ordinary least squares (OLS) regressions on average individual quality-adjusted output, (2) and (5) show random effects (RE) regressions on periodic individual quality-adjusted output, and (3) and (6) show random effects regressed on periodic group quality-adjusted output. The reference category is *ranking*.
When we control for our sample, our results remain unchanged for models (1), (2), (3), (4), and (5) unless for model (6) where the coefficient for peer rating becomes significant (p=0.095).

2.4.2.4 The effect on quality-adjusted output

To assess the robustness of our results, we further analyze the effects of peer evaluation system and feedback transparency on participants' image description quality. Hence we perform regressions on *quality-adjusted*

output $_{Objective}$.[33,34] The reason we use the latter measure is that we know that *quality-adjusted output* is driven by biased ratings (as can be seen from Figure 2.3 biased quality measure. The right panel in Table 2.3 shows ordinary least squares (4), random effects (5), and group clustered random effects (6) specifications on *quality-adjusted output* $_{Objective}$. In columns (4) and (5) we find a significant interaction effect, which suggests that the effect of peer evaluation systems on quality-adjusted output is moderated by the extent to which outcome feedback is transparent. Inspection of the simple effects shows that the effect of transparency is significant in the peer rating conditions (F=4.08, p-value=0.04, un-tabulated) and that the effect of peer evaluation system is significant in the transparent conditions (F=2.86, p-value=0.09, untabulated) for model (4). The simple effects for model (4) likewise show that the effect of transparency is significant in peer rating conditions (β=0.08, p-value=0.09, untabulated) and that the effect of peer evaluation system is significant in transparent conditions (β=0.12, p-value=0.06, untabulated). In addition, we find a significant main effect for the peer evaluation system in models (4) and (5), indicating that when peers evaluate each other by using a peer rating system, both average and individual quality-adjusted output increases significantly compared to when peers evaluate each other using a peer ranking system.

2.4.3 Supplementary analyses

2.4.3.1 The anticipation effect

According to Carpenter et al. (2010) individuals are demotivated to exert effort under peer ratings systems, when the outcomes of such evaluations

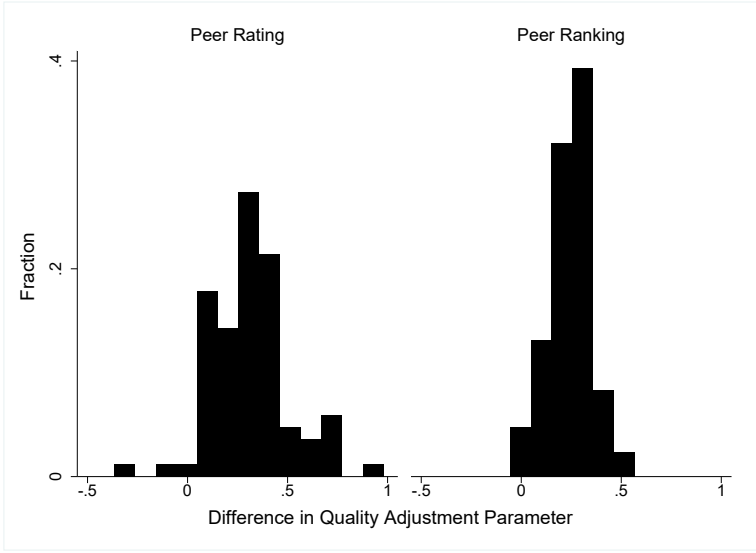
[33] In addition to these variables, we analysed an alternative measure for image description quality, calculated as the number of words in each image description. When analyzing word count as another dependent variable, however we do not find any significant effects (untabulated).

[34] We use our *quality-adjusted output* $_{Objective}$ variable as an alternative quality measure. Prior studies usually construct a measure of creativity/quality using unbiased ratings by having the solutions rated by external raters (Carpenter et al., 2010; Kachelmeier, Reichert, & Williamson, 2008). The reason that we opt for an alternative approach in this paper is simply a practical one. The total number of image descriptions in our study amounted to 6465. To attain reliable ratings of the quality of these image descriptions, we would have needed several raters (Amabile, 1996), who would have had to be compensated accordingly.

are non-transparent, because individuals expect others to underrate them. This is what we refer to as the anticipation effect. To test whether this is indeed the case, we provided the following statement in the post-experimental questionnaire to participants: "I believe other employees assigned bad assessments to increase their own chances of earning more.", measured on a 7 point Likert scale (1=strongly disagree, 7=strongly agree, with 4 as the midpoint). Indeed, participants expected that their group members would underrate them more in the rating condition than in the ranking condition, under non-transparent outcome feedback (4.11 versus 3.55, t_{166}=-2.17, p-value=0.03 two-tailed).

The question then is, are these expectations grounded? Hence, we analyze whether participants indeed underrate each other in non-transparent settings when the peer evaluation system used is peer ratings. We do this by analyzing the difference between *quality-adjustment parameter* $_{Subjective}$ and *quality-adjustment parameter* $_{Objective}$, over all four periods (Carpenter et al., 2010). Figure 2.4 depicts the latter difference when evaluation outcomes are not transparent, for peer ratings and peer rankings. Negative differences reflect underrating behavior and positive differences reflect overrating behavior (i.e. leniency). As can be seen from the left panel in Figure 2.4, only a very small fraction of peer ratings are below the objective quality evaluation (i.e. 2.38%). More interestingly, we observe from Figure 2.4 that evaluations are rather lenient under peer ratings systems when evaluations are not transparent. An insignificant correlation between the *quality-adjustment parameter* $_{Subjective}$ resulting from the peer ratings and the *quality-adjustment parameter* $_{Objective}$ (r=0.06, two-tailed p-value=0.61) shows that the evaluations in the peer rating condition do not correlate with objective evaluations, when these are not made transparent. This provides some evidence of peer rating leniency. Taken together, we argue that the observed effects are consistent with Carpenter et al. (2010) and Balafoutas et al. (2020). Nevertheless, we observe that individuals are rather lenient in their actual rating behavior. This finding suggests that such ratings are likely to be interpreted as underratings in the non-transparent peer rating condition, even though they are fairly lenient.

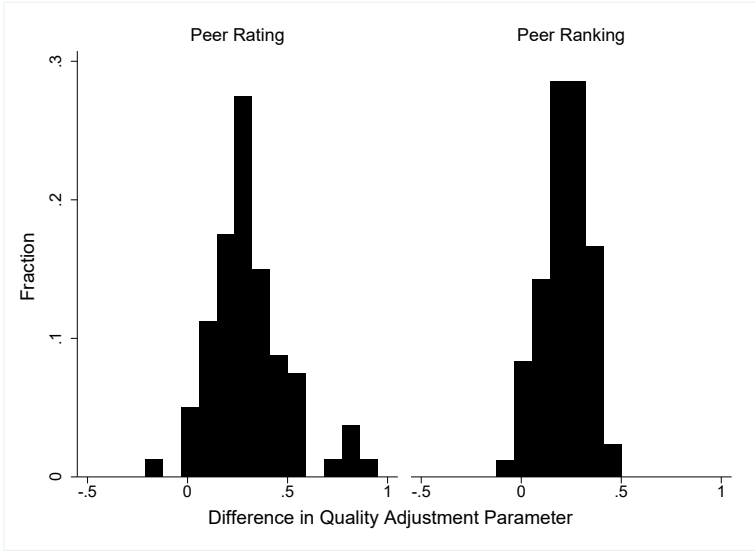
Figure 2.4: Rating behavior in non-transparent conditions



Alternatively, we analyze individuals' rating behavior when peer evaluations are made transparent. We predict that individuals are less likely to underrate peers when evaluations are made transparent because individuals might experience disutility from breaking the social norm of fair evaluations, as the basis of our theory for H1b. Figure 2.5 shows again the difference between *quality-adjustment parameter* $_{Subjective}$ and *quality-adjustment parameter* $_{Objective}$ when peer evaluations are made transparent. As can be seen from the figure, the difference in evaluations is less dispersed for peer rating systems when evaluations are made transparent than when they remain private (see left panels of Figure 4 and Figure 5). Indeed, a significant correlation coefficient between *quality-adjustment parameter* $_{Subjective}$ and *quality-adjustment parameter* $_{Objective}$ ($r=0.26$, two-tailed p-value=0.02) suggests that the evaluations from our study are strongly correlated to the objective evaluations. Additionally, we test whether individuals' peer evaluations correlate with their *quality-adjusted output* $_{Objective}$ in the peer rating conditions. While we find that the peer evaluations, as measured by an individual's rank in the group based on their average received rating/ranking, significantly correlate with *quality-adjusted output* $_{Objective}$ ($r=-0.37$, two-tailed p-value<0.001) when peer evaluations are made transparent, we fail to find significant

correlations between both variables when peer evaluations remain private (r=-0.12, two-tailed p-value=0.29).[35] These correlations thus suggest that the ranks resulting from peer evaluations (1=best evaluated, 4=worst evaluated) are significantly more in line with individual's actual performance in transparent conditions compared to non-transparent conditions. Together, these results provide some support for fairer ratings in transparent conditions, which suggests that individuals are less likely to perceive the ratings as underratings compared to peer ratings in the non-transparent condition.[36]

Figure 2.5: Rating behavior in transparent conditions



Recall that part of our theory for H1b also argues that effort decreases

[35] The reported correlations refer to the first working period. We observe similar correlations for the subsequent work periods (untabulated).

[36] We also analyze whether individuals are evaluated consistently throughout the study. As such, we study whether an individual's rank is the same across all four work periods. To this end, we construct a binary variable indicating one if an individual has the same rank throughout the study and zero otherwise. We find that when peer ratings become transparent, evaluations are significantly more consistent (M=0.200, SD=0.045) than when peer ratings remain private (M=0.083, SD=0.030) (Wilcoxon rank-sum test Z=-2.144; n_1=84; n_2=80; p=0.032 (two-tailed)). For the other conditions, we do not observe differences in evaluation consistency.

when rankings are made transparent compared to when they remain private because they feel that the system is unfair. That is, individuals will not trust the outcomes of the peer evaluations if this is the case. As such, we asked participants whether they agreed with the following statement: "I trusted the outcomes of the peer assessment." on a 7-point Likert scale (1=strongly disagree, 7=strongly agree, with 4 as the midpoint). To examine whether trust in the system differentially affects participants' quality-adjusted output across the *ranking* conditions, we estimate a linear regression equation with quality-adjusted output $_{\text{Objective}}$ as the dependent variable and the following independent variables: a dummy taking the value "1" ("0") if the outcomes from the peer ranking were made transparent (non-transparent), our measure of trust in the system[37], and the interaction between the former two. The results of this analysis shows a positive significant interaction effect, suggesting that the trust in the system has a stronger effect on quality-adjusted output $_{\text{Objective}}$ when the peer rankings are made transparent relative to when they remain private (p-value=0.046 two-tailed, untabulated). A further inspection of the marginal effects shows that trust in the system is lower in the transparent condition, which in turn is associated with lower quality-adjusted output $_{\text{Objective}}$.

Alternatively, we examine whether individuals might coordinate their rating behavior in the condition where peer evaluations are made transparent, by giving co-workers consistently high ratings. Table 2.1, Panel A shows the descriptive statistic for the mean average received ratings across our conditions. We compute the average of these ratings across all four working periods. Individuals in the transparent peer rating condition do not significantly rate each other higher (M=5.64, SD=0.13) than individuals in the non-transparent condition (M=5.58, SD=0.14), as indicated by the results of a t-test (t_{162}=-0.28, two-tailed p-value=0.78). Additionally, we test whether the distribution of the ratings in the transparent and non-transparent peer rating conditions are significantly different from each other. Again, we find no differences as indicated by the chi-squared test (X^2_{55}=50.70, two-tailed p-value=0.64). Together, these results show no evidence of collusion between participants in the transparent peer rating condition.[38]

[37] We first mean-centered this measure.

[38] To provide some more support for our theory, we examine whether participants indeed responded in a fairer way across transparent conditions. The post-task

Finally, we examine whether participants in the ranking conditions indeed felt uncomfortable performing the peer assessment whenever the quality of the image descriptions differed little across peers. We asked all participants whether participants agreed to the following statement: "When the quality of the descriptions of my colleagues did not differ sufficiently, I felt uncomfortable assessing their output quality.", measured on a 7 pint Likert scale (1= strongly agree, 7=strongly disagree, with 4 as the midpoint). Indeed, participants in the ranking conditions were more uncomfortable assessing their peers when this was the case, compare to participants in the rating conditions (4.43 versus 3.92; t_{330}=2.85; p-value<0.01 two-tailed).

2.5 Conclusion and discussion

While prior literature has shown that peer evaluations can be detrimental to employee effort due to an anticipation effect where individuals expect their peers to under-evaluate them (Balafoutas et al., 2020; Carpenter et al., 2010), our theory and results suggest that these negative effects (in settings where the outcomes of such peer evaluations are not transparent) can be mitigated either by installing peer rankings or by making the peer ratings transparent. This study finds that when individuals work in a self-managing group, individual efforts on a crowdsourcing task are higher under a peer ranking scheme when its outcomes are kept private rather than a peer rating scheme. We find that the opposite relation holds when peer evaluation outcomes are transparent.

Building on prior accounting research, our findings suggest that the effects of forced rating systems used by supervisors (Bol, Kramer, & Maas, 2016; Cardinaels & Feichter, 2021; Moers, 2005), can be extended to peer evaluation settings. That is, similar to the above studies, we find that performance deteriorates when individuals are asked to evaluate their peers using a forced rating system, but only when its outcomes are made transparent.

---

questionnaire provided participants with the following statement: "I assessed my peers' work quality honestly.". Participants responded using a 7-point Likert scale (1=strongly disagree, 7= strongly agree, with 4 as the midpoint). Results show that participants in the transparent conditions indeed tend to assign more honest peer evaluations compared to participants in the non-transparent conditions (5.98 versus 5.70; t_{330}=-2.19; p-value=0.03 two-tailed).

Similar to Evans and colleagues (2016) and Maas and Van Rinsum (2013) we show that ratings become more accurate as transparency increases when individuals are asked to evaluate their co-workers with a peer rating system. However, we do not observe this pattern for peer ranking systems as we find that evaluations are in line with actual performance for most periods (untabulated).

As with any experimental study, this study has its limitations. First, this study was run online, which could potentially interact with our results. Prior studies have found that the mere physical presence of peers in a room can affect employee effort and performance (Falk & Ichino, 2006; Mas & Moretti, 2009). Therefore, it could be interesting to examine whether the effects found in this study, hold in a non-online setting. Second, our study is characterized by a highly competitive setting (i.e. highest quality-adjusted performer from the group wins a bonus). Future research might explore the potential boundary conditions that could help further our knowledge of the effects of peer evaluation systems. More specifically, future research could shed light on whether different types of incentive schemes (such as fixed wage or piece-rate schemes) interact with peer evaluation design. Given that participants in this study are confronted with an effort allocation decision towards quantity and quality, future research could examine variations in how both output dimensions are incentivized and how this affects employee effort (Rubin et al., 2018). This study makes use of a real-effort experimental task, however, prior research suggests incentives might affect effort differently when studies make use of chosen effort tasks (Brüggen, & Strobel, 2007). Hence, future research could examine whether our results would extend to chosen effort settings. Relatedly, future research examining the effects of peer evaluation design and its use by a supervisor on employee effort is instructive (Arnold et al., 2018, 2020). The set-up of the current paper closely resembles a crowdsourcing setting where individuals work in self-managing groups, without a controlling supervisor. Individuals might evaluate each other differently when a supervisor has to decide on their final compensation for example, as this might induce image concerns (Ariely, Bracha, & Meier, 2009). Together with this, future research could examine whether other types of transparency, such as process transparency (rather than outcome transparency) affect employee effort in the presence of an evaluating supervisor, through a procedural justice lens for example (Colquitt, 2012;

Thibaut and Walker, 1975). Alternatively, as employees in practice are typically not only evaluated by their peers but also by their supervisors, it might be interesting for future research to examine whether employees react differently to peer evaluations compared to supervisor evaluations. Prior research has already established that knowledge sharing between employees with status differences behaves differently as compared to knowledge sharing between equal-status employees (Haesebrouck, Cools, & Van den Abbeele, 2018). As such, future research could examine the effects of status differences on employee effort in a performance evaluation context. Third, the use of the labels in our peer evaluation systems (below average, average, above average) might have introduced a relative component into our peer *rating* condition. That is, participants might have used those labels as anchors to compare peer output quality to each other, which could have introduced a researcher expectancy effect (Cook et al., 2002). Hence, future research might examine whether peer evaluation labels might interact with other design characteristics. Finally, one might argue that the use of forced rating systems is not appropriate for small group sizes, as individuals might feel disappointed with their outcome. Therefore, future research could examine whether the effects of peer evaluation design interact with group size.

Chapter 3

Narrative comments in peer evaluations: evidence on individual acceptance levels and team creativity*

Abstract:

In this study, we examine how the use of narrative comments in peer evaluations and their purpose (monetary versus non-monetary) influences employee acceptance levels of the peer evaluation system in an experiment where participants perform a creative task in a team. We predict and find that the inclusion of narrative comments in peer evaluations increases employee acceptance levels more when the evaluations are used by managers to determine bonuses (monetary purpose) than when managers only provide feedback messages (non-monetary purpose). However, we find preliminary evidence that individuals facing peer evaluations including narrative comments are more likely to focus on their personal goals rather than the group goals in the monetary purpose condition compared to the non-monetary condition as shown by the decrease in creative performance over time, because of increased impression management motives. Taken together, we contribute to the literature by showing that the effects of peer narrative comments on employee outcomes depend on the use of these evaluations by managers and that peer evaluations with a monetary purpose can have unintended consequences for team creativity.

* This chapter is co-authored with Eddy Cardinaels and Alexandra Van den Abbeele. We thank Markus Arnold, Sophie De Winne, Christoph Feichter, and Sabra Khajehnejad. We also thank Daniel Alejo and Sjuul Derkx for their research assistance. Tilburg University's TiSEM Institutional Review Board, and KU Leuven's Social and Societal Ethics Committee both approved this study.

3.1 Introduction

Firms are increasingly making use of so-called peer-to-peer recognition tools, in which employees can recognize each other's contributions, thereby for example saving points that can later be redeemed for prizes.[39] Other forms of peer-to-peer recognition tools offer employees the possibility to give real-time praise, typically through a public social platform (Arnold, Hannan, & Tafkov, 2018; Mosley, 2015). These peer recognition tools are often used in cooperative settings to better assess individual contributions, but the format of these peer-to-peer recognition systems can vary strongly. While some systems serve as a public social platform (i.e. LinkedIn), other systems work as point systems (i.e. Nectar, Bonusly), or as rating systems (i.e. Workhuman). Moreover, the purpose of those peer-to-peer recognition tools differs to the extent that they are being used as input for compensation or in other organizations more for developmental purposes (Appelo, 2015; Brutus, 2010).

Although there is evidence on the positive consequences of such systems in terms of employee retention, satisfaction and productivity (SHRM, 2018), little research has examined which factors determine whether employees will accept and consequently use these systems (Maley, Dabic, & Moeller, 2020). We follow Maley et al. (2020) by referring to acceptability as the extent to which employees engage with, use, and consider peer-to-peer evaluation systems as an added value. That is, the concept goes beyond the notion of feedback acceptance, which is defined by prior research as the employee's belief that the outcomes of peer-to-peer evaluation systems are an accurate portrayal of their performance (Ilgen, Fisher, & Taylor, 1979, p. 356; Loftus & Tanlu; 2018, p. 280). The level of acceptability for these evaluation systems is a potential driver for employee behavior (Ilgen, Fisher, & Taylor, 1979; Ilgen, & Davis, 2000). For instance, Amazon's peer-to-peer recognition tool received criticism as employees felt uncomfortable using the system (Arnold et al., 2018). Likewise, Cappelli and Tavis (2016) show that influential organizations such as Google and Deloitte also abandon their performance evaluation systems as they lack employee acceptability. While acceptability of peer-

[39] Examples of firms using these types of recognition tools are "The Motley Fool" (Tiny Pulse, 2020), "Zappos" (Zappos Insights) and "NASA" (Gallus, Jung, & Lakhani, 2019).

to-peer systems is an important condition of employee involvement and engagement which might subsequently affect employee outcomes, not many studies have looked into the factors affecting the acceptability of peer-to-peer systems.

The goal of this study is twofold. First, we examine how the level of acceptability of peer evaluation systems is affected by two design characteristics, namely their format and the purpose for which they are being used. We consider a monetary purpose of peer evaluations on the one hand, and a non-monetary peer evaluation purpose on the other hand. When a peer evaluation's purpose is monetary, firms typically tie the outcomes of peer evaluations to compensation, while in a non-monetary purpose peer evaluation outcomes are not tied to a bonus. However, prior research on the use of (peer) evaluation systems and their purposes reports mixed results. While some authors find detrimental effects on employee behavior of the use of peer evaluations tied to compensation (Bamberger, Erev, Kimmel, & Oref-Chen, 2005; Bettenhausen & Fedor, 1997; Carpenter, Matthews, & Schirm, 2010; Fedor, Bettenhausen, & Davis, 1999; Tavoletti, Stephens, & Dong, 2019), others report the opposite (Arnold et al., 2018; Glover & Xue, 2020; Kelly, Dinovitzer, Gunz, & Gunz, 2020). We also consider the format of the peer evaluation, and more specifically the role of narrative comments. While many studies use rating scales (Balafoutas, Czermak, Eulerich, & Fornwagner, 2020; Carpenter et al., 2010; Dewaele, Cardinaels, & Van den Abbeele, 2022; Leibbrandt, Wang, & Foo, 2018), fewer studies examine narrative comments in peer evaluations (Lampe, Shäffer, & Schaupp, 2021). Nevertheless, the majority of U.S. companies, surveyed by Gorman, Meriac, Roch, Ray, & Gamble (2017) make use of these narrative comments. Prior research argues that the use of narrative comments versus quantitative methods (i.e. ratings, rankings, point systems) can affect employee behavior differently (Brutus & Donia, 2010; David, 2013; Bentley 2019). It is thus interesting, both from a theoretical as well as a practical point of view, to examine the effects of narrative comments.

To understand how peer evaluation purpose and format affect acceptability levels, we develop our prediction based on impression management theory. More specifically, we expect the format of peer evaluations to moderate the effect of the peer evaluation purpose on acceptance levels because of impression management concerns. That is, acceptance levels might be higher when narrative comments are added to

the evaluation when the its purpose is monetary, but not when the purpose is non-monetary. We predict that individuals have image concerns (Ariely, Bracha, & Meier, 2009), which are larger when individuals can receive a possible bonus. Consequently, under monetary purpose peer evaluations individuals will try to create a good impression of themselves through these peer evaluations, in order to get higher chances of earning a bonus. We argue that individuals can use narrative comments in peer evaluations to engage in impression management by behaving altruistically. That is, adding the possibility of narratives in a monetary purpose peer evaluation might lead to higher levels of acceptance, as individuals are nicer towards each other. Alternatively, impression management tactics such as behaving altruistically and being nice to each other, are likely to be mitigated under non-monetary purpose peer evaluations.

The second goal of this study is to explore whether and how feedback resulting from peer evaluations impacts team creativity. Prior research establishes that individuals learn from obtained feedback resulting from performance evaluation systems, which can in turn positively affect their creative performance (Son & Kim, 2016; Joo, Song, Lim, &Yoon, 2012). While prior literature argues that higher levels of acceptability might increase the likelihood of integrating peer feedback into a cooperative task (like creativity) (Ilgen et al. 1979, Ilgen & Davis, 2000), we conjecture that it could have unintended consequences on creative performance. We formulate a research question based on insights from both the feedback proactivity literature (De Stobbeleir, Ashford, & Buyebs, 2011; Joo, et al 2012) and management accounting literature relating to the surrogation phenomenon (Bently, 2019; Choi, Hecht, & Tayler, 2012, 2013). We expect that when narrative comments are included in monetary purpose peer evaluations, individuals will focus more on performing well on a specific measure (managing impressions through their peer evalutions) which could cause them to perform lower on the creative team task (Bentley, 2019; Choi, Hecht, & Tayler, 2012, 2013). Alternatively, we expect the inclusion of narrative comments to increase team creative performance more when the peer evaluation purpose is non-monetary than when it is monetary. We argue that in this setting, impression management concerns are lower compared to when peer evaluations are used for monetary purposes.

To test our predictions, we conduct an online interactive experiment on Amazon's Mechanical Turk (MTurk). We use a 2x2 between-subjects design in which we manipulate peer evaluation purpose (monetary versus non-monetary) and peer evaluation format (inclusion of narrative comments versus not). The peer evaluation is constructed such that workers have a total of 100 points that they can assign to their peers, either with or without narrative comments. Workers are allocated to groups of four, in which three are assigned the role of employee and one the role of manager. Employees then work on a creative task for three consecutive periods. We use a creative context, as employees are able to observe their co-workers' actions while working on the task. In practice, managers lack the resources to assess each employee's contribution to a group output (Brun & Dugas, 2002), and thus such a setting makes the use of peer feedback relevant. Additionally, managers might find it difficult to evaluate their team's output because of its subjective nature (Cardinaels, Dierynck, & Hu, 2020). Often, when individuals work in a group setting, creativity can arise through collaborative and cooperative behavior from group members (Adler, & Chen, 2011; Toubia, 2006). Hence, managers might rely on peer evaluations in order to gather more information on group dynamics. In between each period, employees are required to perform a peer evaluation, by evaluating their peers' contributions to the development of a creative (group) proposal. The outcomes of these peer evaluations are then sent to the manager, who can use them to either decide on a bonus allocation (monetary purpose) or feedback message allocation (non-monetary purpose).

In line with our expectations, we find that acceptance levels are highest when peer evaluations include narrative comments (in addition to points) and the purpose is monetary. We find support for our theory that individuals in this condition indeed engage more in impression management than individuals in the other conditions. We observe several forms of impression management tactics, as participants use more integrative negotiation tactics (Essa, Dekker, & Groot, 2018; Giebels, De Dreu, & Van De Vliert, 2000; Van den Abbeele, Roodhooft, & Warlop, 2009), use less swear words and a more positive tone in their narrative comments (Bell & Arthur, 2008; Brett & Atwater, 2001). However, we do not observe that teams in this condition adapt their creative performance over time, suggesting that the feedback from peer evaluations is not well incorporated in future creative endeavors. That is,

we do not find statistical evidence for the mediating effect of acceptance on team creativity. More interestingly, we show that the use of narrative comments in non-monetary purpose peer evaluations does improve team creative performance over time compared to monetary purpose peer evaluations.

With this study, we contribute to the literature in a number of ways. First, we expand the understanding of narrative feedback in performance evaluations. Prior literature studying narrative feedback as a management control tool remains scarce (Arnold, Ponick, &Schenk-Mathes, 2008; Arnold et al., 2018; Brutus, 2010; David, 2013; Lampe et al., 2021; Stubbs, 2021). In our study, we show that the inclusion of narrative comments in peer evaluations can be useful but simultaneously not always effective for firm performance. We do find that employees report higher levels of acceptance when peer evaluations include narrative comments. Yet, we also show that team member communication based on peer evaluations does not always affect team (creative) performance in a positive sense. Specifically, while prior literature suggests acceptability is an important factor for positive behavioral responses (Ilgen et al. 1979, Ilgen & Davis, 2000, Lampe et al., 2021), we show that such acceptance in the context of monetary purpose peer evaluations does not help the firm to reach higher levels of creative performance. As such, we add to the management accounting literature by showing that when managerial discretion determines bonus allocations (Arnold et al., 2018; Arnold, Hannan, & Tafkov, 2020), narrative feedback may increase system acceptability, but it does not help people to spark sufficient motivation for future creativity.

We also add to prior accounting literature on creativity. Several accounting scholars have studied the effects of various individual incentives on creativity (Brüggen, Feichter, & Williamson, 2018; Cardinaels et al., 2020; Chen, Williamson, & Zhou, 2012; Kachelmeier, Bernhard, & Williamson, 2008). The current literature, however, remains scarce with respect to the effects of evaluation systems by superiors on creativity, with Cardinaels & Feichter (2021) as a notable exception. We add to this stream of literature by showing that the design of peer performance evaluation systems can significantly affect team creative performance. We also extend this research stream to team creativity (Chen et al., 2012) instead of prior accounting research that predominantly focuses on the role of management control systems that

affect individual creativity (Cardinaels, & Feichter, 2021; Cardinaels et al., 2021, Kachelmeier et al., 2008).

3.2 Background and hypothesis development

3.2.1 Background

In this study, we aim to increase our understanding of the acceptability of peer evaluation systems. We follow Maley et al. (2020) who refer to the acceptability of a performance evaluation system as the extent to which employees engage, use, and consider the evaluation system as an added value. While feedback acceptance contributes to system acceptability (Maley et al., 2020), the concept of system acceptability goes beyond feedback acceptance. Hence, factors affecting feedback acceptance, are predicted to ultimately also affect system acceptability.

The extent to which individuals accept feedback from performance evaluations, depends on a number of factors. Prior evidence shows that high-quality feedback increases trust perceptions (Coletti, Sedatole, & Towry, 2005; Moers, 2005), and more recent findings indicate that feedback quality further increases acceptance levels (Son & Kim, 2016). These increased trust perceptions then (as a predictor of feedback acceptance), have been found to affect cooperative behavior positively in a number of studies examining the effect of control precision (Anderson, Cheng, & Phua, 2021; Christ, Sedatole, & Towry, 2012; Christ, Sedatole, Towry, & Thomas, 2008; Coletti et al., 2005; Hofmann, Lei, & Grant, 2009). Additionally, prior research identifies feedback favorability as another determinant of feedback acceptance (Ilgen et al., 1979). As such, several scholars find that individuals tend to accept feedback more when it is positive or favorable than when it is negative (Anseel & Lievens, 2006, 2009; Bell & Arthur, 2008; Brett & Atwater, 2001; Stone & Stone, 1985; Tonidandel, Quiñones & Adams, 2002).

3.2.1.1 Peer evaluation purpose: monetary versus non-monetary

We use insights from economics and psychology to examine the role of narrative comments in peer evaluations under different tournament incentives.[40] In this study, we incorporate managerial discretion in (non-

[40] According to Gürtler and Harbring (2010) tournaments can take various forms, including the contest for monetary bonuses or other non-monetary benefits. Non-

monetary) bonus allocations in a context where it is difficult for managers to learn what each member contributes to the group output. We, therefore, use a creative task where individuals work together in a team, such that they have more information about their group functioning than their manager. Managers then ask peers to evaluate each other and make their final decision based on peer evaluations they receive from employees. However, the purpose for which peer evaluations are used differs across organizations (Gorman et al., 2017). Prior research shows that the extent to which peer evaluations are used for monetary or non-monetary purposes by management determines how employees react to it. Early research by Fedor and colleagues (1999) argues that peer evaluations are generally more accepted by employees when they are being used for non-monetary or more training-like purposes, while acceptability is lower when peer evaluations are used for monetary purposes (pay or promotion decisions). Likewise, Hecht, Hobson, and Wang (2020) find that frequent performance evaluations have a negative effect on task performance, especially when individuals know that the evaluations will be used for a specific purpose (i.e. use of the performance report for evaluating *critical reasoning skill* by the administrator). Similarly, a number of studies in organizational sciences find that when (peer) performance evaluations are used for monetary purposes rather than non-monetary purposes, negative effects on employee performance can occur (Bamberger, Erev, Kimmel, & Oref-Chen, 2005; Bettenhausen & Fedor, 1997; Fedor & Bettenhausen, 1989; Fedor et al., 1999; Tavoletti et al., 2019). This finding has also been established in economics, where Carpenter and colleagues (2010) find reduced effort levels when peer evaluations are used to calculate compensations.

While the above studies provide evidence on the negative effects of using peer evaluations for monetary purposes, some studies do suggest that peer evaluations can work when the outcome is tied to compensation. For example, a study by Arnold and colleagues (2018) finds that the use of team subjective communication (which is a form of peer evaluation)

monetary benefits can be thought of as relative performance information (such as provided by recognition programs), which has been showed to motivate individuals in tournament settings (Hannan, Krishnan, & Newman, 2008; Tafkov, 2013; Wang, 2017). Hence, both the purpose peer evaluations in this study constitute tournament incentive schemes.

increases performance in a bonus pool allocation setting. Likewise, a field study examining profit allocations among law firm partners finds that when the performance evaluation systems include more subjective measures of performance, partners allocate the profits in a fairer way (i.e. they accede less to the wishes of clients and fellow partners) (Kelly et al., 2020). In a similar vein, Glover and Xue (2020) find that principals take into account more subjective measures of team member performance when allocating the bonus pool when they have discretion on the size of the bonus pool. This then in turn increases team performance under a team incentive rather than an individual incentive scheme (Glover & Xue, 2020).

These prior studies propose mixed evidence with respect to the purpose of peer evaluations on various outcome variables. Evidence on the factors that determine peer evaluation acceptability is lacking (Maley et al., 2020) and may explain why evidence is mixed. That is, if a system is accepted it might render more positive outcomes. As such, we examine how (peer) evaluation system characteristics can affect why peer evaluations are accepted and ultimately alter can employee behavior.

3.2.2 Hypothesis development

3.2.2.1 The interactive effect of narrative comments and peer evaluation purpose on acceptability

We argue that the format of peer evaluation moderates the effect of peer evaluation purpose on acceptance levels. Specifically, we propose that peer evaluations induce higher levels of acceptability when the format of the evaluation includes narrative comments in a monetary purpose peer evaluation rather than a non-monetary purpose peer evaluation. The inclusion of narrative comments in a peer evaluation may serve as an impression management tool. Individuals are namely motivated by others' perceptions (Ariely et al., 2009). That is, individuals have the desire to be liked and approved by others and to avoid creating negative impressions of themselves. As Ariely and colleagues (2009) argue, these impression motivations can lead to altruistic behavior.

The extent to which employees engage in impression management depends on the possibility to gain favorable evaluations from supervisors (Nadler, Ellis, & Bar, 2003). As managers in our setting base their evaluations on the outcomes from peer evaluations - thereby having full

discretion - employees face substantial uncertainty regarding this subjectivity in their final evaluations (Bol, 2008; Gibbs, Merchant, Stede, & Vargus, 2004; Prendergast, 1999). As such, individuals are motivated to make a good impression on their manager, through these peer evaluations (Bol, 2008).

Individuals can create a favorable impression of themselves by giving favorable feedback to peers. We expect individuals to utilize narrative comments in peer evaluations as a means to be liked by peers. Prior research has documented that individuals engage in 'ingratiation' as an impression management strategy, which is defined as an attempt to be liked by others, by flattering their peers (Drory & Zaidman, 2007). This flattery then is predicted to result in peer evaluations with a positive or favorable tone. Therefore, we expect individuals to write positive comments about each other as a way of creating a positive impression as negative comments are less likely to contribute to one's impression positively. Given that prior literature finds that positive feedback is perceived as more credible by recipients (Brett & Atwater, 2001; Stone & Stone, 1985), we expect that this will increase acceptability levels among individuals (Bell & Arthur, 2008). Moreover, we believe that individuals facing a monetary purpose peer evaluation rather than a non-monetary purpose peer evaluation will engage in impression behavior to a larger extent, in particular when individuals can use narrative comments to evaluate each other. Under a monetary purpose evaluation system, employees are motivated to avoid creating a negative impression as this might result in not being allocated a bonus at all, while employees in non-monetary purpose evaluations incur a lower cost of creating negative impressions. We conjecture that unfavorable events in terms of 'losing a bonus' will outweigh unfavorable events in terms of 'a negative feedback message'. That is, we expect that peer evaluations induce greater image motivation when there is money at stake which can ultimately lead to more positive peer evaluations. This positivity then can affect the extent to which feedback is accepted.

Alternatively, when peer evaluations are used for non-monetary purposes, individuals are expected to learn from performance feedback. Individuals might therefore give more honest or accurate peer feedback in the narrative comments. While accurate feedback might determine higher acceptability, the extent to which the feedback is negative can also decrease acceptance levels (Bell & Arthur, 2008). Hence, we expect that

the use of narrative comments in non-monetary purpose peer evaluations are less likely to increase acceptance levels.

When peer evaluations do not include narrative comments, individuals are asked to allocate points to each other. These quantitative type of peer evaluations, make social comparison among group members more salient as they tend to focus on the total allocated points they receive (Brutus, 2010). We, therefore, expect individuals in both the monetary and non-monetary purpose peer evaluations to compete for points. This competition, in turn, might both increase productive and unproductive efforts (Wang, 2017). As such, dishonesty in peer evaluations might be a concern for individuals (Carpenter et al., 2010), ultimately affecting their fairness perceptions of the peer evaluation system (Arnold et al., 2018; Chan, & Thornock, 2022). Hence, we expect individuals to perceive points-only peer evaluations as less acceptable compared to peer evaluations that include narrative comments. Taken together, we state our first hypothesis as follows:

Hypothesis 1: *Peer evaluation purpose and format interact such that acceptance levels are highest when monetary purpose peer evaluations include narrative comments, relative to non-monetary purpose peer evaluations including narrative comments.*

3.2.2.2 Effect of peer evaluation purpose and format on creative performance: exploratory thoughts

In this study, we examine the effect of peer evaluation purpose and format in a creative task setting in which individuals can observe each other's contributions to the creative group output, while managers do not have access to this type of information. Although prior research advocates that high feedback acceptance, as part of evaluation system acceptability, leads individuals to integrate and act upon the received feedback (Ilgen & Davis, 2000; Ilgen, Fisher, & Taylor, 1979), even in creative tasks (Son & Kim, 2016), we argue that it might not always have the intended effects on creative performance when peer evaluations are used for monetary purposes. Recall that we predict that impression management motives play an important role in peer evaluations when their purpose is monetary. More specifically, as we expect that peer evaluations will be more positive when they contain narrative comments which impact feedback acceptance positively, it might lack constructive

feedback needed for improving one's creative performance. Indeed, recent findings show that impression management motives might impede individuals from seeking help related to creative tasks, which can decrease creative performance altogether (Carnevale, Huang, Vincent, Farmer, & Wang, 2021).

When peer evaluations are used for non-monetary purposes, we expect that impression management concerns are less prevalent. This might in turn open up the focus on the team goal, namely developing creative group proposals rather than the individual goal (getting a high evaluation from the manager). As such, we expect individuals to use the narrative comments in peer evaluations as means to provide critical information to their peers that can improve overall team creativity, rather than to engage in flattery. This type of information, critical or challenging feedback, could in turn increase creative performance as creativity usually results from the ability to think out-of-the-box, and the ability to combine insights from different approaches (Amabile, 1996). Individuals thus learn from such feedback which in turn increases their creative performance (Son & Kim, 2016; Joo, Song, Lim, & Yoon, 2012). Additionally, prior research argues that individuals become more creative when they obtain diverse input of feedback (De Stobbeleir et al., 2011; Madjar, 2005; Perry-Smith & Shalley, 2003).

Alternatively, when peer evaluations are used for monetary purposes and include narrative comments we expect individuals to focus too much on getting allocated the highest bonus by their manager by creating a positive impression of themselves. This is in line with the surrogation phenomenon (Choi, Hecht, & Tayler, 2012, 2013), where scholars posit that individuals focus too much on performing well on a specific measure on which they are being compensated, thereby losing sight of what is really important (i.e. in our setting team creativity). Indeed Cardinaels and Feichter (2021) find in their paper that when managers use forced ratings to evaluate creative ideas, individuals' performance decreases because of the worry about the evaluation criteria. In addition, prior evidence also shows that creative efforts (operationalized as the discovery of production efficiencies) are lower when compensation is based upon targets (Webb, Williamson, & Zhang, 2013). Our theory development above leads us to expect that the inclusion of narrative comments and their effect on team creativity depends on peer evaluation purpose. Taken together, we pose the following research question:

Research question: *Does the combination of points and narrative comments increase team creativity more when the peer evaluation purpose is monetary than when it is non-monetary?*

3.3 Experimental method and design

To test our hypotheses, we conduct a 2x2 between-subjects experiment on Amazon's Mechanical Turk (MTurk)[41] in which we manipulate the purpose of the peer evaluation (i.e. monetary versus non-monetary) and the format of the peer evaluation system (i.e. points versus points combined with narrative comments). The experiment is coded using oTree, a Python-based framework for conducting online interactive experiments (Chen, Schonger, & Wickens, 2016).

3.3.1 Participants

In total, 373 workers completed the study. We retained data from 315 workers for further analyses, excluding observations from 58 workers based on the following exclusion criteria. First, we exclude 35 workers who either submit only numbers, parts of the instructions, blank responses, or insults during the three consecutive idea development phases. Because our experiment involves a task in which employees should contribute to a joint group output, by inputting responses to open questions in the idea generation phase, employees are required to put reasonable effort into the task. Hence, these invalid responses of 35 workers are excluded as these workers, may have not provided their best effort (Bentley, 2021; Clor-Proell, Guggenmos, & Rennekamp, 2020). Second, given that we run our experiment online, we further exclude 23 workers who failed one or more attention checks. That is, we provided participants in our post-experimental questionnaire with the following statements: "If you read this, indicate that you strongly disagree.", and "If you read this, indicate that you strongly agree." Workers responded using a 7-point Likert scale (1=strongly disagree, and 7=strongly agree). Workers' age ranges from 21 to 65, with a mean age of 38.28 (SD=10.33

[41] Participants on MTurk are pre-screened, by using a minimum HIT approval rate of 98%, minimum number of accepted HITs of 500, and a US location. By default, all MTurk workers are at least 18 years old, as Amazon requires workers to be at least 18 years old when signing up for an MTurk account.

years), and 54.92% are male.[42] On average, workers earned $6,21 to participate in a study taking 53 minutes of their time.

3.3.2 Experimental task

3.3.2.1 Employee task

MTurk workers are first matched into four-person groups. In each group, one participant is assigned the role of manager, and the other three participants are assigned the role of employee. The group composition and participant roles remain unchanged throughout the course of the experiment. After participants read the instructions and complete a comprehension quiz, the experimental task starts.

The employees in each group work on a creative solution for a business problem for three consecutive periods, each lasting 5 minutes (adapted from (Cardinaels et al., 2020; Cardinaels & Feichter, 2021; Chen et al., 2012)). We adapt our procedures from Chen and colleagues (2012) such that employees are able to generate and develop creative ideas online, on which group members can decide to build further on. More specifically, employees have two options. They either decide to submit a parent idea, or they decide to build further on an existing parent idea, by commenting a child idea (see Appendix 3.1). This task allows for different ways of working and contributing which would allow employees to either offer a rating about the function of their colleagues (points only condition) or to describe something about their colleagues' functioning (narrative condition), that could be informative to the manager who ultimately evaluates his/her employees with either a bonus or a message (purpose manipulation).

Employees worked on the creative task for three consecutive periods and are presented with the following business problems: how to (1) "cut costs at airline companies", (2) "ensure a good work/life balance for people that are working from home", (3) "help reduce climate change at the office". Each business problem is only shown to the employees at the start of its respective round. The goal of the experimental task is for employees to cooperate in developing a creative solution to each business problem. We define a creative solution as a solution that is "original, innovative, and implementable within a reasonable budget", in line with

[42] From all 315 workers, only one worker did not want to disclose his/her gender.

previous studies on creativity (Amabile, 1996; Cardinaels et al., 2020; Chen et al., 2012).

At the end of the 5-minute periods, each group gets 2 minutes to decide which idea gets submitted to the manager, using a chat function (similar to the design of Cardinaels et al. 2020). During this 2-minute chat, each employee can select the number of the parent idea he/she wants to submit to their manager. This submitting procedure is essentially a voting system, which allows each employee the same amount of *power* in deciding which creative proposal is submitted to the manager.[43] After the creative proposal is submitted to the manager, employees perform their peer evaluations while they see the output of the idea generation phase.

3.3.2.2 Manager task

The manager's main task is to provide each of his/her employees with a final evaluation, based on their peer evaluations. During the time that the employees are generating creative ideas, the manager reads an article that describes each business problem. After the employees submit their creative proposal, the manager's screen is automatically updated such that he/she can read the creative proposal while the employees perform their peer evaluations. After employees have performed their peer evaluations, the outcomes are automatically sent to the manager.

Depending on the assigned condition, the manager either uses the peer evaluation to determine each of his/her team member's compensation (monetary purpose) or to give a written feedback message to each employee (non-monetary purpose) (see Appendix 3.2, Panel A, and Panel B). Managers in our experiment have full discretion in allocating their evaluations. That is, managers can decide to give each employee the highest evaluation, but also differentiate their evaluations across employees.[44]

[43] The creative proposal that will be submitted to the manager is either the one that is selected by the majority of employees in the group (i.e. by two employees), or otherwise the program selects a random number from the submitted numbers of all employees in the group.

[44] We opt for this design choice as forced rating systems have been showed to undermine creativity (Cardinaels & Feichter, 2021) by for example introducing intragroup competition (Chen et al., 2012).

3.3.3 Manipulations

3.3.3.1 Format manipulation

We first manipulate the format of the peer evaluations between participants at two levels (inclusion of narrative comments or not). All employees are instructed to assess their fellow employees' contribution toward the creative solution development, in a peer evaluation. Recall that individuals in the experimental group task can either choose to submit a parent idea or to contribute to an already existing idea by submitting a child idea. Contributing to the creative solution then can be done through both submitting parent ideas and child ideas. However, participants were free in using the available information from the idea development phases for the evaluation of their peers' contributions. To this end, employees in all conditions are informed to allocate points towards their fellow employees. Each employee has a total of 100 points to give away, and he/she can distribute these in any way they like. Our format manipulation then consists of including narrative comments in the peer evaluation in addition to the point allocations. More specifically, employees in the narrative comments conditions are asked to provide comments concerning their fellow employees' contribution to the creative solution. Moreover, the instructions state that "these comments may include areas that the peer in question needs to work on, recommendations for improvement, etc." These narrative comments thus provide peers with feedback, that can enhance their contributions to the team (De Stobbeleir, Ashford, & Zhang, 2020). All group members can review the outcomes of the peer evaluations after each idea generation phase, along with their manager's final evaluation (on the same screen).

3.3.3.2 Purpose manipulation and payoffs

Additionally, we manipulate the purpose of the peer evaluation between participants at two levels. Similar to prior literature on (peer) performance evaluations, we distinguish between monetary and non-monetary purposes (Brutus, 2010; Fedor & Bettenhausen, 1989). This manipulation is implemented by instructing the participants assigned the role of managers, to use the peer evaluations in order to evaluate each employees' contribution towards the creative proposal. Managers in the monetary purpose conditions are instructed to evaluate employees' contributions as below average with a $0 bonus, average with a $3 bonus,

and above average with a $ 6 bonus. Alternatively, managers in the non-monetary purpose conditions are instructed to evaluate employees' contributions with a message (see Appendix 3.2).[45,46] Additionally, we instruct participants in the (non-)monetary purpose conditions that their manager will decide on their final evaluation, based on the outcomes of the peer evaluations by allocating employees a(n feedback message) bonus.

This manipulation also determines participant compensation. Each participant earns a fixed participation fee of $3 and has the possibility to earn a bonus on top of that. More specifically, employees assigned to the monetary purpose conditions, are informed that their bonus is determined based on the random selection of one period. We opt for this design choice to ensure that workers are motivated during all three idea development phases, similar to prior research (Hannan, Towry, & Zhang, 2013). This means participants can either earn $3, $6, or $9 in total (depending on their managers' final bonus allocations). Alternatively, employees in non-monetary purpose conditions are informed that they earn a fixed $3 bonus on top of their participation fee, implying a total fixed compensation of $6. Likewise, managers in all conditions are informed that their final compensation amounts to $6.

3.3.4 Dependent variables

3.3.4.1 Acceptability levels

We construct a three-item scale to measure the extent to which extent employees accept the peer evaluation system. We adapt our (post-experimental) items from Fedor and Bettenhausen (1989), and based on

[45] The reason that we opt for individual incentives in this team context is that prior research shows that organizations often make use of individual incentives when creativity is valued (Chen et al., 2012).

[46] One might argue that the use of feedback messages can induce strong intragroup competition because of its relative component. However, recall that managers have full discretion w.r.t. the allocation of these messages, making the condition comparable to the monetary purpose condition. Alternatively, prior research suggests that tournament incentives in the form of non-monetary creativity peer rankings is effective in incentivizing creativity (Charness, & Grieco, 2019). As such, Charness and Grieco (2019) find that individual's creativity increases with their rank, by putting more effort in the creative task, suggesting that individuals learn from these peer ranks and change their behaviour accordingly.

definitions of acceptance in prior literature (Ilgen et al., 1979; Loftus & Tanlu, 2018, Maley et al., 2020). First, participants indicated the extent to which they agreed to whether the "peer evaluation was a waste of time and effort", on a 1 (strongly disagree) to 7 (strongly agree) scale (reverse-coded). Second, participants indicated the extent to which "the peer evaluation added value to the creative solution development", using the same 7-point Likert scale. Finally, the participants indicated the extent to which they thought "the peer evaluations were useful", on the same 7-point Likert scale.

Exploratory factor analyses are conducted to investigate the extent to which the former three items explain the same factor. To this end, the three items are subjected to a principal component factor analysis with varimax rotation (see Table 3.1). Based on Kaiser's eigenvalue greater than 1 rule of thumb, one factor is retained (eigenvalue = 2.070), indicating that all three items load on the same underlying factor. The Chronbach's alpha is equal to 0.754, indicating that the scale has good internal consistency. We create our measure of employee acceptance by averaging employees' responses to the three items (as reported in Table 3.2).

Table 3.1: Factor analysis on employee acceptability

Three-item scale (eigenvalue = 2.070; explained variance = 0.690)					
Item	**Loading**	**Mean**	**Min**	**Max**	**Std. Dev.**
The peer evaluation was a waste of time and effort (reverse-coded)	0.656	4.903	1	7	1.908
I think the peer evaluation added value to the creative proposal development	0.925	4.899	1	7	1.633
The peer evaluations I received were useful	0.886	4.907	1	7	1.751

All items are measured using a 7 point Likert scale, ranging from 1=strongly disagree, to 7=strongly agree.

3.3.4.2 Change in creativity

As prior literature prescribes, we invite independent raters to assess the creativity of the submitted (group) proposals (Amabile, 1996). As such, we recruit eight different students from the online participant pool of the university's research lab, and give them €20 as compensation. These

students rate all submitted proposals on a scale from 0 (not very creative) to 100 (very creative) (Brüggen, Feichter, & Williamson, 2018; Cardinaels & Feichter, 2021; Kachelmeier, Reichert, & Williamson, 2008). All raters are presented the same instructions. After reading the instructions, raters complete a short quiz to ensure they understand the definition of creativity. Similar to Cardinaels and Feichter (2021) raters take a short break after assessing the submitted proposals to one of the three business problems. That is, raters assess the proposals to business problems in the same order as they are presented to employees in the MTurk experiment. The Chronbach's alpha for the proposals for the first business problem is 0.71, 0.82 for proposals for the second business problem, and 0.79 for proposals for the third business problem. These values are all well above the reliability thresholds, indicating good interrater reliability (Peterson, 1994). We construct our measure of creativity by averaging creativity ratings of all independent raters. Since we are interested in examining how creativity changes after feedback is given, we construct a change measure by calculating the difference in creativity from the first to the third idea generation phase.

3.4 Results

3.4.1 Checks

Before reporting our results, we first test whether the random assignment of workers to our conditions was successful. To this end, we estimate a multiple linear regression to test whether workers' individual characteristics collectively affect being allocated to one of the experimental condition. Results show no significant joint effect ($F_{5,309}$=1.45; p-value=0.21; Adjusted R-squared=0.01). Likewise, workers did not differ with respect to their individual characteristics on being allocated the role of manager or employee ($X^2_{5,315}$=3.85; p-value=0.57; Pseudo R-squared=0.01).

Second, we test whether our manipulations were successful. In the post-experimental questionnaire we ask employees to indicate the extent to which they agree to "when performing the peer evaluations, I did this by writing comments" on a 7-point Likert scale (1=strongly disagree, 7=strongly agree). The mean response for this item in the points only condition (M=3.86, SD=2.14) was lower than for the narrative comments

condition (M=6.03, SD=0.83). The difference between both means is significantly different from zero (t_{235}=10.30, p-value<0.001 two-tailed). Additionally, we ask managers to indicate the extent to which they agree to "the peer evaluations received, included point allocations only", using the same 7-point Likert scale. The mean response for managers in the points only condition was higher (M=6.00, SD=1.38) than the mean response for managers in the narrative comments condition (M=3.92, SD=2.32). Again, the difference between both means is significantly different from zero (t_{76}=-4.81, p-value<0.001 two-tailed). Together, these results suggest that our peer evaluation format manipulation was successful. Finally, we ask employees whether their bonus was allocated by their manager, using the same 7-point Likert scale. The mean response on this item was higher (M=6.25, SD= 1.06) for employees in the monetary purpose conditions than for employees in the non-monetary purpose conditions (M=5.64, SD=1.56). This difference is significantly different from zero (t_{235}=-3.53, p-value<0.001 two-tailed). Likewise, we ask managers whether they "could allocate monetary bonuses to their employees", using the same 7-point Likert scale. The mean response on the latter item was again higher for managers in the monetary purpose conditions (M=6.55, SD=0.85) than managers in the non-monetary purpose conditions (M=4.11, SD=2.18), and the difference is again significantly different from zero (t_{76}=-6.60, p-value<0.001 two-tailed). Together, these results suggest that our peer evaluation purpose manipulation was successful.

3.4.2 Test of hypothesis

The descriptive statistics for employee acceptability are shown in Table 3.2 and graphed in Figure 3.1. We predict that narrative comments have a different influence on employee acceptability, depending on the peer evaluation purpose. The level of acceptability across conditions is different and its pattern is consistent with our hypothesis. The highest acceptance levels are observed in the monetary purpose condition when peer evaluations include narrative comments (M=5.31, SD=1.31). To test our first hypothesis we first run an analysis of variance (ANOVA). Panel A of Table 3.3 presents the results. The results report a significant main effect of peer evaluation format (F=5.22, p-value=0.02 two-tailed). However, we fail to find a significant interaction effect between peer evaluation purpose and format on employee acceptability (F= 1.66, p-

value=0.20 two-tailed). The significant main effect indicates that the use of narrative comments in peer evaluations significantly increases employee acceptability, compared to when peer evaluations consist only of allocated points.

We then use contrast coding as an additional test for our hypothesis (Buckless & Ravenscroft, 1990; Guggenmos, Piercey, & Agoglia, 2018; Rosenthal & Rosnow, 1985). We assign contrast weights of +3 to the condition in which peer evaluations include narrative comments and are used for monetary purposes, and -1 to the three remaining conditions. The results of our contrast test are presented in Table 3.3, Panel B. The planned contrast test is significant (F=6.35, p-value=0.01 two-tailed). This significant effect supports our hypothesis that employee acceptability is highest when peer evaluations include narrative comments, which are then subsequently used by managers to determine bonus allocations (monetary purpose). The simple effects also confirm that format (i.e. adding narratives) has an effect in the monetary purpose condition (F=6.47, p-value=0.01 two-tailed) but not when peer evaluations are used for non-monetary purposes (F=0.49, p = 0.49), which is consistent with our prediction.

Collectively, these results suggest that employee acceptability of peer evaluation systems can be improved by including narrative comments, but only when these peer evaluations are used by managers to determine bonus allocations. Contrary to what Brutus (2010) proposed (p. 153), we find that not the quantitative information, but the qualitative information (i.e. narrative comments) influences employee reactions (i.e. acceptance levels) greater when peer evaluations are used for monetary purposes (i.e. determining bonus allocations).

		Monetary Purpose		Non-Monetary Purpose	
		Points	**Points + Narrative Comments**	**Points**	**Points + Narrative Comments**
Panel A: Dependent Variables					
Acceptability		4.645 (1.542)	5.311 (1.306)	4.737 (1.459)	4.922 (1.416)
Number of observations Individual level		61	59	57	60
Creativity	Period 1	40.665 (16.116)	42.399 (16.719)	43.500 (10.408)	40.775 (13.736)
	Period 2	44.057 (20.860)	45.536 (15.167)	42.950 (15.912)	38.156 (18.537)
	Period 3	46.188 (18.063)	39.554 (17.808)	44.094 (15.589)	50.188 (12.060)
	Change	5.523 (23.822)	-2.845 (18.072)	0.594 (15.974)	9.413 (19.073)
Number of observations Group level		22	21	20	20
Panel B: Process Variables					
Impression management		5.508 (0.1512)	5.847 (1.172)	5.772 (1.180)	5.517 (1.347)
Affective responses		3.415 (1.898)	2.791 (1.515)	3.538 (1.639)	3.644 (1.834)
Content narrative comments	Negative emotions	N.A.	0.223 (1.026)	N.A.	2.070 (9.043)
	Swear words	N.A.	0.000 (0.000)	N.A.	0.486 (2.668)
	Tone	N.A.	96.415 (10.868)	N.A.	88.699 (27.423)
Number of observations Individual level		61	59	57	60
Negotiation tactics	Integrative tactics	9.182 (4.043)	11.00 (5.683)	10.650 (4.209)	10.400 (4.903)
	Distributive tactics	2.500 (2.596)	2.286 (1.901)	1.350 (1.387)	1.650 (1.872)
Difference (integrative – distributive) tactics		6.682 (5.158)	8.714 (6.157)	9.300 (4.015)	8.750 (6.157)
Number of parent ideas developed	Period 1	5.864 (2.816)	7.428 (3.414)	6.650 (3.558)	6.700 (3.197)
	Period 2	5.955 (2.681)	8.809 (5.046)	6.850 (3.717)	7.300 (4.426)
	Period 3	6.545 (2.703)	9.286 (5.226)	6.800 (4.162)	8.550 (6.436)
	Mean	6.121 (2.191)	8.508 (4.339)	6.767 (3.466)	7.517 (4.254)
Number of observations Group level		22	21	20	20

This table presents the mean, (standard deviation) for the variables, for each condition.

Figure 3.1: Observed pattern of results for employee acceptability

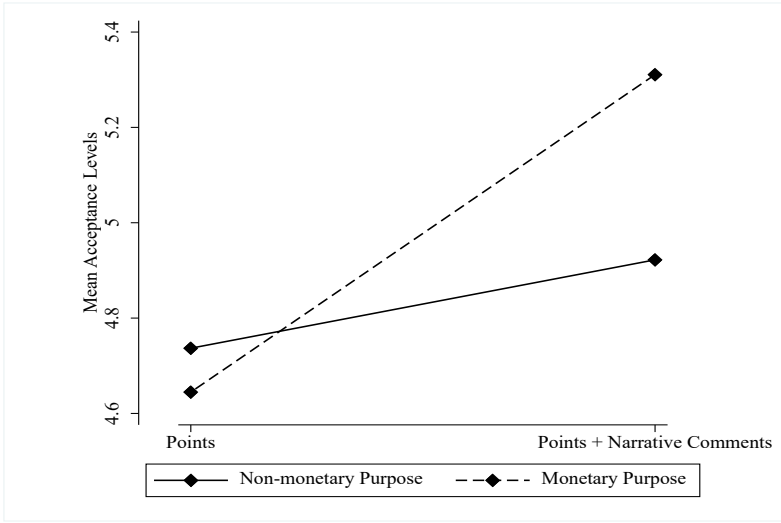


Table 3.3: ANOVA on acceptability

Panel A: ANOVA results				
	df	**M.S.**	**F**	**p-value**
Narratives	1	10.728	5.22	0.023**
Monetary Purpose	1	1.301	0.63	0.427
Format x Purpose	1	3.418	1.66	0.199
Residual	233	2.056		
Panel B: Contrast test				
	Contrast		**F**	**p-value**
	1.628		6.35	0.012**
Panel C: Follow-up simple effects				
	df		**F**	**p-value**
Effect of *Format* in the Monetary purpose condition	1		6.47	0.012**
Effect of *Format* in the Non-Monetary purpose condition	1		0.49	0.485
Effect of *Purpose* in points only condition	1		0.12	0.728
Effect of *Purpose* in narrative comments condition	1		2.18	0.141

***, **, * Indicate p-values at $p < 0.01$, $p < 0.05$, and $p < 0.10$, respectively.
All p-values are reported two-tailed. Adjusted R-squared = 0.0188.

3.4.2.1 Test of theory: impression management motives

We base our prediction of H1 on impression management theory. Recall that we expect individuals to engage more in impression management when peer evaluation purposes are monetary, especially when these include narrative comments. Individuals can use impression management tactics by 1) being nice to each other during the idea generation phases (by chatting nicely to each other) and 2) being nice to each other using narrative comments as a means to give favorable feedback to peers. To test for our theory, we conduct a series of statistical tests.

Similar as Towry (2003) we analyze chat messages between employees to construct a measure of cooperation. While Towry (2003) does not analyze the content of communication between participants, literature on negotiation tactics suggests that the discussion between individuals determines whether they engage in cooperation or non-cooperation (De Dreu & McCusker, 1997; Giebels, De Dreu, & Van De Vliert, 2000; Kelley & Stahelski, 1970). Negotiation has been defined as the discussion between individuals relying on each other aiming at reaching an agreement (Pruitt, 1981). As such we analyze the content of the chat messages by coding all messages based on an adapted coding scheme (see Appendix 3.3) on team communication, similar to prior studies (Essa et al., 2018; Giebels et al., 2000; Van den Abbeele et al., 2009).[47] The coding scheme was developed by the principal investigator, based on theories of negotiation tactics (Giebels et al., 2000; Pruitt, 1981; Van den Abbeele et al., 2009). Then, the codes were applied to the chats by two (external) coders who were blind to the experimental manipulations (Krippendorff, 2018). Their inter-coder agreement attained acceptable levels (Krippendorf $_{Cu}\alpha$=0.86). The chat messages are coded and analyzed in ATLAS.ti.

[47] The communication between group members in our study can be thought of as a negotiation, where individuals enter a discussion on *which* proposal group members would like to submit to their manager. Individuals thus, have the opportunity to put forward suggestions and engage in a discussion with the goal of reaching a consensus or agreement (Pruitt, 1981). Nevertheless, we do acknowledge that this form of communication can be thought of a *softer* form of negotiation where individuals are not likely to experience direct costs or losses from negotiation outcomes, in our setting. Therefore we only focus on the concepts of negotiation communication relevant to our study (see Appendix 3.3).

As shown by prior research, negotiators (i.e. individuals aiming to reach an agreement) can either use integrative or distributive tactics. Integrative tactics are tactics aimed at problem-solving, an invitation for cooperation, and creating a pleasant atmosphere (Chang, Cheng, & Trotman, 2013; Essa et al., 2018; Van den Abbeele et al., 2009). Distributive tactics, on the other hand, involve tactics using persuasive arguments, withholding information, and creating unpleasant atmospheres, which could eventually result in conflicting discussions (Essa et al., 2018; Graham, Evenko, & Rajan, 1992; Pruitt & Lewis, 1975).

First, we conduct a contrast analysis on the difference in negotiation tactics (Buckless & Ravenscroft, 1990; Guggenmos et al., 2018; Rosenthal & Rosnow, 1985), of which the descriptives are presented in Table 3.2. Positive differences in negotiation tactics reflect more cooperative behavior (i.e. using more 'integrative tactics' while negative differences reflect less cooperative behavior (i.e. using more 'distributive tactics') (Essa et al., 2018). We again assign a weight of +3 to the condition in which peer evaluations include narrative comments and are used for bonus allocation purposes. The remaining three conditions are assigned a weight of -1. Untabulated results indicate that the contrast is significant (F=2.98, p-value = 0.09 two-tailed). This result thus shows that teams in the monetary purpose conditions where peer evaluations include narrative comments, use more integrative tactics during the 2-minute chat (over all idea generation phases), compared to teams in other conditions. This further suggests that indeed, teams make the effort to create a pleasant and cooperative working atmosphere, with the goal to receive favorable peer evaluations.

Second, we analyze the content of narrative comments. Similar to Lampe et al. (2021) we use the Linguistic Inquiry and Word Count (LIWC) dictionary to analyze the content of the comments (Pennebaker, Boyd, Jordan, & Blackburn, 2015). Untabulated results show that the content of narrative comments is more negative in the non-monetary purpose condition, as compared to the monetary purpose condition (t_{117}=1.56, one-tailed p-value=0.06). Moreover, the narrative comments in the non-monetary purpose condition contain more swear words than the comments in the monetary purpose conditions (t_{117}=1.40, one-tailed p-value=0.08). Together, these findings suggest that the content of narratives is more positive and nicer in monetary purpose peer

evaluations, indicating that individuals indeed engage in impression management motives.[48] These results are also in line with prior literature suggesting that positive feedback is more likely to be accepted by individuals (Bell & Arthur, 2008; Brett & Atwater, 2001).

Third, we provide further evidence on individuals' tendency to engage in impression management motives, by analyzing how participants answered the question "Because the manager decided on my final evaluation, I was eager to make a good impression.", measured on a 7-point Likert scale (1=strongly disagree, 7=strongly agree). Results from an ANOVA analysis on the latter item as dependent variable, by including both the main effects of our independent variables (peer evaluation purpose and format), and their interaction. The results are depicted in Table 3.4, and show a significant interaction effect (F=3.03, two-tailed p-value=0.08). This result suggests that individuals in monetary purpose peer evaluations including narrative comments indeed engage in impression management to a greater extent compared to when peer evaluations are used for non-monetary purposes only. Nevertheless, the follow-up simple effects show no significant effects.

[48] Some examples of positive narrative comments include: "wonderful supportive ideas for the idea that was submitted", "great ideas, love that they both submitted their own and built upon others", some examples of negative narrative comments include: "Poor cooperation", "terrible", "this employee doesn't have good points with him".

Table 3.4: ANOVA on impression management

Panel A: ANOVA results				
	df	**M.S.**	**F**	**p-value**
Narratives	1	0.104	0.06	0.806
Monetary Purpose	1	0.067	0.04	0.844
Format x Purpose	1	5.232	3.03	0.083*
Residual	236	1.725		
Panel B: Follow-up Simple Effects				
		df	**F**	**p-value**
Effect of $Format$ in the Monetary purpose condition		1	1.10	0.294
Effect of $Format$ in the Non-Monetary purpose condition		1	2.00	0.159
Effect of $Purpose$ in points only condition		1	1.19	0.277
Effect of $Purpose$ in narrative comments condition		1	1.89	0.171

***, **, * Indicate p-values at $p < 0.01$, $p < 0.05$, and $p < 0.10$, respectively.
All p-values are reported two-tailed. Adjusted R-squared = 0.0006.

3.4.3 Research question

Our research question relates to the expected effects of our manipulations (purpose and format) on team creative performance. The descriptive statistics regarding the creativity of teams' submitted proposals are shown in Table 3.2. Recall that participants in our experiment receive feedback after each idea generation phase. Hence, participants have the possibility to adapt their behavior accordingly. To this end, we analyze the change in creative performance from period 1 to period 3. As shown in Table 3.2, the pattern of changes provides some preliminary support for our research question.

We run an ANOVA on the change in creative performance. The results are reported in Table 3.5 and show that the format and purpose of peer evaluations significantly interact (F=4.00, p-value = 0.05 two-tailed) on the change in creative performance. Further inspection of the follow-up simple effects shows that the effect of peer evaluation purpose is significant in peer evaluations including narrative comments (F=4.03, p-value=0.05 two-tailed). These results indicate that individuals integrate the feedback from peer evaluations to a higher extent when they include

narrative comments and the purpose is non-monetary, rather than monetary. These findings suggest that indeed, higher employee acceptability can increase the integration of feedback which ultimately leads individuals to act upon the feedback and improve their creative performance, but only when the peer evaluation purpose is non-monetary. Interestingly, the provision of narrative comments in monetary purpose peer evaluations results in worse creative performance (M=-2.85, SD=18.07) compared to when only points are allocated in peer evaluations (M=5.52, SD=23.82). This difference fails to be significant, however, the one-tailed p-value of 0.10 is close to the conventional cut-off levels. Together, these findings provide some initial rationale for failing to find a significant mediating effect of employee acceptability on change in creative performance (untabulated).

Table 3.5: ANOVA on change in creative performance

Panel A: ANOVA results				
	df	**M.S.**	**F**	**p-value**
Narratives	1	1.053	0.00	0.958
Monetary Purpose	1	278.188	0.73	0.396
Format x Purpose	1	1529.900	4.00	0.049**
Residual	79	382.396		
Panel B: Follow-up Simple Effects				
		df	**F**	**p-value**
Effect of $Format$ in the Monetary purpose condition		1	1.97	0.164
Effect of $Format$ in the Non-Monetary purpose condition		1	2.03	0.158
Effect of $Purpose$ in points only condition		1	0.67	0.417
Effect of $Purpose$ in narrative comments condition		1	4.03	0.048**

***, **, * Indicate p-values at $p < 0.01$, $p < 0.05$, and $p < 0.10$, respectively.
All p-values are reported two-tailed. Adjusted R-squared = 0.0202.

3.4.4 Supplementary analyses

3.4.4.1 Effect of personal goals on team creativity – an analysis of the generation of unique parent ideas

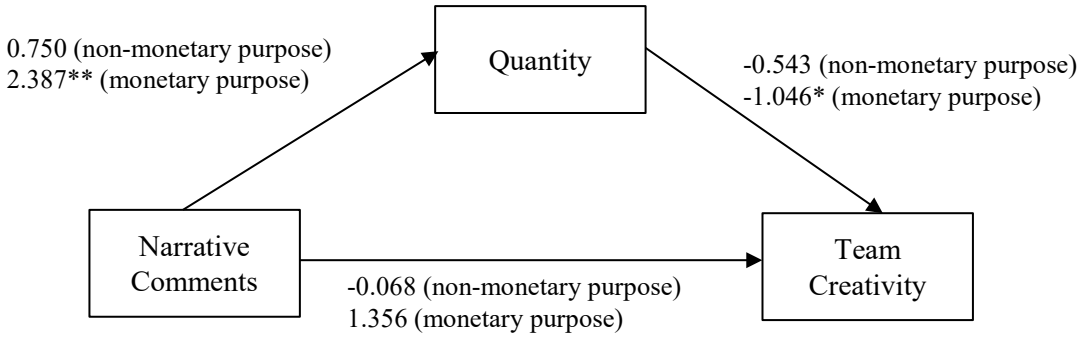
Prior research suggests that individuals who focus on their personal goals (i.e. creating a favorable image of themselves) rather than on group goals (i.e. developing a creative proposal) might fail to adopt the group goals

as their personal goals (Ellemers, De Gilder, & Haslam, 2004; Towry, 2003; Van Knippenberg, 2000). Indeed, a number of studies have found that independent individual efforts towards creativity do not extend to group creativity settings (Chen et al., 2012; Kachelmeier et al., 2008; Kachelmeier & Williamson, 2010). As such, similar to Chen et al. (2012) we explore whether groups in the monetary purpose conditions tend to work more independently by analyzing the number of initial parent ideas generated by each group. That is, we calculate the mean number of parent ideas generated during all three idea development phases per group, which constitutes our measure of quantity. We then estimate a mediation model to test whether the volume of generated parent ideas mediates the relationship between narrative comments and team creativity, following Hayes (2018, Model 4).

The results of our mediation model are shown in Figure 3.2. The mediation model is estimated using the bootstrap method with 1,000 replications and bias-corrected 90% confidence intervals (Hayes, 2018). Under non-monetary purpose peer evaluations, the inclusion of narrative comments does not affect the average number of parent ideas generated relative to peer evaluations without narrative comments (0.75; 90% CI [-1.26, 2.76]). And the number of generated parent ideas under developmental purpose peer evaluations does not affect the average creativity of the submitted proposals (-0.543; 90% CI [-1.15, 0.07]). More interestingly, we do find that the inclusion of narrative comments in monetary purpose peer evaluations increases the average number of parent ideas generated (2.39; 90% CI [0.72, 4.06]) which in turn decreases average team creativity (-1.05; 90% CI [-2.03, -0.06]). However, the indirect effect of narrative comments on team creativity through the quantity of generated ideas is not significant (-2.50; 90% CI [-5.36, 0.37]). Collectively, these findings provide some initial evidence for the loss of creativity of group proposals when individuals are dealing with monetary purpose peer evaluations, due to their focus on personal goals rather than their group goals. We argue that individuals focus on generating a high volume of parent ideas as a strategy to manage impressions (which could eventually affect their peer evaluations) as a personal goal. This personal goal then does not contribute to the group goal of creating a creative group proposal, which is usually a process of individuals exchanging ideas and contributing to each other's ideas (De Stobbeleir et al., 2020).

Figure 3.2: Path model

Panel A: Mediation model



Panel B: Estimated effects – non-monetary purpose				
	Effect	**Bootstrapped SE**	**Lower 90% CI**	**Upper 90% CI**
Indirect effect	-0.407	0.670	-1.510	0.695
Total effect	-0.475	3.158	-5.670	4.720
Panel C: Estimated effects – monetary purpose				
	Effect	**Bootstrapped SE**	**Lower 90% CI**	**Upper 90% CI**
Indirect effect	-2.497	1.742	-5.362	0.368
Total effect	-1.140	3.921	-7.590	5.310

Panel A presents the proposed mediation model graphically for participants in both peer evaluation purpose conditions (monetary and non-monetary), showing the direct effects (standardized coefficients) of the presented variables. Narrative Comments is defined as a binary variable indicating 1 when peer evaluations include narrative comments and 0 otherwise. Quantity is measured as the average number of parent ideas created over all periods, in a given team. Creativity is measured as the average creativity of submitted group proposals over all periods.
Panel B presents the bootstrapped estimations following Hayes (2018) with bias-corrected confidence intervals, for participants in de non-monetary purpose conditions.
Panel C presents the bootstrapped estimations following Hayes (2018) with bias-corrected confidence intervals, for participants in de monetary purpose conditions.
***, **, * Indicate two-tailed p-values at $p < 0.01$, $p < 0.05$, and $p < 0.10$, respectively.

3.4.4.2 Why does team creativity increase when narrative comments are used in non-monetary purpose peer evaluations?

We further explore the mechanisms through which group creativity increases in non-monetary purpose peer evaluations. First, we examine

whether individuals in these conditions collaborate to a greater extent by contributing to their group members' ideas. To this end, we follow Chen and colleagues (2012) by analyzing the depth of idea development. Our measure of idea development depth is similar to Chen et al. (2012) as we count the number of child ideas from each submitted proposal. Nevertheless, since our setting is characterized by three idea generation periods, we calculate the change in the number of child ideas from period one to period 3. The higher the change of idea development depth, the more individuals contribute to each other's ideas over the course of our experiment. We find no significant differences for change in idea development depth when the peer evaluation format includes narrative comments as compared to when they do not (0.65 vs. 0.40, t_{38}=-0.48, two-tailed p-value>0.10). Hence, we cannot conclude that individuals in non-monetary purpose conditions collaborate more when narrative comments are present in those peer evaluations than when they are not.

Second, we examine whether the tone of the narrative comments might increase creativity. Prior research has established that the tone of mood states can significantly affect creative performance (De Dreu, Baas, & Nijstad, 2008). More specifically, De Dreu and colleagues (2008) find that moods with positive and negative tones can both increase creative performance, although the effect depends on the extent to which individuals find themselves in an activated mood (i.e. moods in which individuals have higher cognitive perseverance and persistence than deactivated moods). Later findings indicate that the extent to which positive group affective tone (i.e. affective reactions within a group) has a beneficial effect on team creativity, depends on the extent to which team members trust each other (Tsai, Chi, Grandey, & Fung, 2012). If trust between team members is low, Tsai and colleagues (2012) find that a positive group affective tone decreases team creativity. Alternatively, prior findings indicate that negative group affective tone can increase team performance outcomes through increased team task conflict, and learning (Chi & Lam, 2021; Jordan, Lawrence, & Troth, 2006). The aforementioned studies thus suggest that the tone of group interactions is predictive of team outcomes, and more specifically team creativity.

To this end, we explore whether the content or tone of the narrative comments affects creative performance differently when peer evaluation purposes are non-monetary rather than monetary. Prior research has established that positive and negative group affective tones have direct

consequences on group interactions (Cole, Walter, & Bruch, 2008; Weiss & Cropanzano, 1996). Hence we assume that stronger affective responses are associated with narrative comments less positive in tone.[49] As such, we compute the average tone (based on the LIWC dictionary developed by Pennebaker et al. 2015) of all narrative comments for each group. Higher values of tone indicate a more positive tone. Descriptive statistics show that the tone of the narrative comments is more positive in peer evaluations that have a monetary purpose compared to a non-monetary purpose (M=96.19, SD=10.87 vs. M=88.70, SD=27.42). To study whether tone impacts creative performance, we perform a linear regression on the difference in creativity with tone and peer evaluation purpose as independent variables (including their main effects and their interaction effect). The results are depicted in Table 3.6 and show that peer evaluation purpose and the tone of narrative comments significantly interact (-0.19, two-tailed p-value=0.04). Figure 3.3 graphs this interaction using predictive margins from the former linear regression model. The figure depicts the predicted change in creativity on the y-axis for different levels of tone in narrative comments on the x-axis. The marginal effects suggest that team creativity decreases more as the tone of narrative comments becomes more positive, under monetary purpose peer evaluations vs. non-monetary purpose peer evaluations (all two-tailed p-values ≤ 0.09, untabulated).

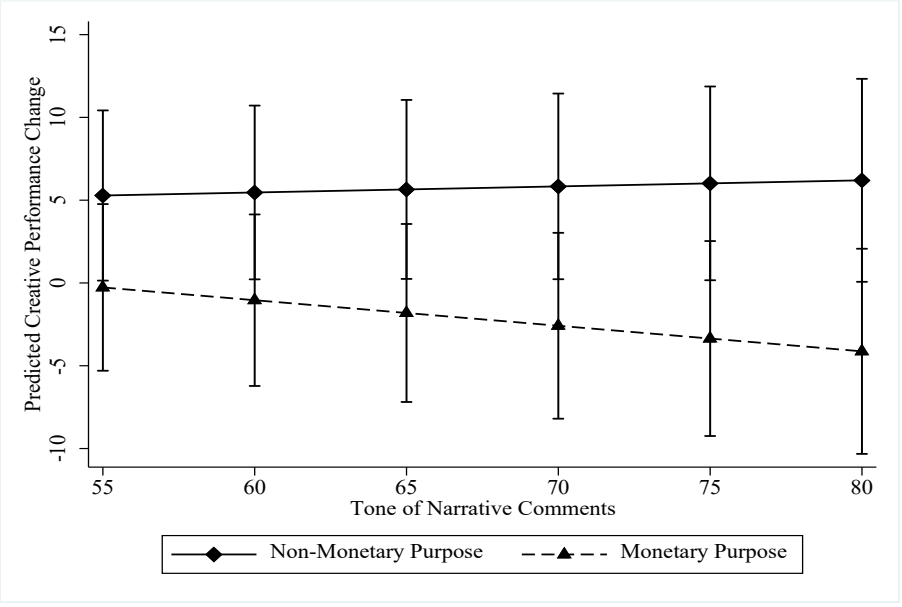
[49] We note that unfortunately we did not capture individuals' affective responses right after they interacted with their group members (and before performing peer evaluations). Future research is thus needed to test whether this assumption would hold.

Table 3.6: Linear regression on change in creative performance

	Change in creative performance
Tone	0.037
	(0.063
Monetary Purpose	4.956
	(5.889)
Tone x Monetary Purpose	-0.191**
	(0.089)
Constant	3.265
	(4.277)
N	83
Adjusted R^2	0.047

This table shows the results of regressing the *Change in creative performance* on *Tone* (a continuous variable where high values indicate a positive tone in narrative comments), *Monetary Purpose* (where 1 indicates peer evaluations used for monetary purposes and 0 otherwise), and their interaction term.
***, **, * indicate two-tailed p-values at $p < 0.01$, $p < 0.05$, and $p < 0.10$, respectively.

Figure 3.3: Predicted creative performance change based on tone and peer evaluation purpose



This figure depicts the predicted changes in creativity as a function of tone levels in narrative comments, across peer evaluation purpose conditions, using 90% confidence intervals..

Third, we examine whether affective responses to peer evaluation feedback might affect changes in team creative performance. We measure affective responses in our post-experimental questionnaire by asking participants to indicate the extent to which they felt 1) very unpleasant, 2) very stressful, and 3) highly aroused when receiving the peer evaluations (Kluger & DeNisi, 1996), all measured on a 7-point Likert scale (1=strongly disagree and 7=strongly agree).[50] The difference between individuals' affective responses in the monetary (M=2.79, SD=0.20) and non-monetary (M=3.64, SD=0.24) peer evaluation purpose conditions that include narrative comments, is significantly different from zero (t_{117}=2.77, two-tailed p-value < 0.01). That is, individuals in the non-monetary purpose conditions react stronger to peer evaluation feedback including narrative comments, than individuals in the monetary purpose condition. We argue that affective responses are impacted by the tone of narrative comments, as shown by prior literature (Loftus & Tanlu, 2018; Bell & Arthur, 2008; Kluger & DeNisi, 1996). Indeed, our measure of affective responses negatively correlates with narrative comment tone (ρ =-0.14, two-tailed p-value = 0.13), however, this correlation is not significant. Nevertheless, this provides us with some limited evidence that stronger affective responses are associated with narrative comments less positive in tone. This limited evidence is in line with Loftus and Tanlu (2018) who find that negative feedback indeed increases affective responses, which in turn increases changes in performance. While we do not find statistical evidence of a mediating effect of affective responses on narrative comments and changes in team creative performance, we do suspect that these affective responses can help explain the higher levels of team creativity in our non-monetary purpose condition. However, future research is needed to further validate this relationship.

Finally, we perform another test to analyze whether responses to receiving feedback might help in understanding how the effect of narrative comments in peer evaluations affect team creativity, depending on the purpose of those evaluations. In our post-experimental

[50] A confirmatory factor analysis on the three items used to measure affective responses, resulted in a one-factor solution after a varimax rotation. Untabulated results of this factor analysis suggest that all three items measure the same underlying construct. The affective response scale is constructed by averaging the responses of all three items, and results in an internally reliable scale (α=0.861).

questionnaire, we ask participants whether agree with the following statement "I changed my behavior in the next period based on the peer evaluations I received in the previous period", on a 7-point Likert scale (1=strongly agree, and 7=strongly disagree). Responses to this statement were higher in peer evaluations including narrative comments when used for non-monetary purposes (M=4.50, SD=0.24) rather than for monetary purposes (M=4.25, SD=0.25). However, the difference between both means is small and not significant (t_{117}=0.72, one-tailed p-value > 0.10). Therefore we cannot conclude that individuals act upon the feedback received from narrative comments in peer evaluations. We do suspect that narrative comments in peer evaluations can alter behavior in team creative contexts and that the effect depends on the purpose for which these peer evaluations are being used.

Altogether, these results suggest that more positive tones in narrative comments have a detrimental effect on team creativity when peer evaluations are used for bonus allocation purposes. Alternatively, we show that the use of narrative comments can be beneficial in increasing team creativity when peer evaluations are used for non-monetary purposes and that the tone of these narrative comments can further increase team creativity. Hence, future research examining the role of narrative comments and their effect on subsequent team creativity is needed to further validate this conclusion.

3.5 Conclusion and discussion

Firms are increasingly making use of peer evaluations to evaluate employee performance in teams. These peer evaluations differ greatly with respect to their formats. And while some firms use these evaluations to determine bonuses, others use them for more developmental or training purposes. Hence, this study examines whether the format and purposes of these peer evaluations affect employee outcomes.

The results of our experiment show that the inclusion of narrative comments in peer evaluations increases employee acceptability of the evaluation system when managers use these as input for their bonus allocation decisions. Our study further presents exploratory evidence on the effects of peer evaluations as feedback mechanisms on creative team performance. We observe that team creativity increases when peer evaluations include narrative comments and their purpose is non-

monetary, while team creativity decreases when peer evaluations are used for monetary purposes.

Our results suggest that narrative comments in peer evaluations can have beneficial effects on employee outcomes. Nevertheless, we find that the inclusion of narrative comments can also have unintended effects on team creativity when managers use these to determine bonus allocation. This finding is in line with recent evidence indicating that individuals want to maintain a certain creativity image among their peers and therefore not engage in creative help-seeking behavior (Carnevale et al., 2021). Hence, we argue that while individuals do report higher acceptability when peer evaluations include narrative comments and are used for monetary purposes, they are too concerned about their image compared to their peers which deters them from actually integrating the feedback resulting from those peer evaluations. This then in turn, negatively affects team creativity.

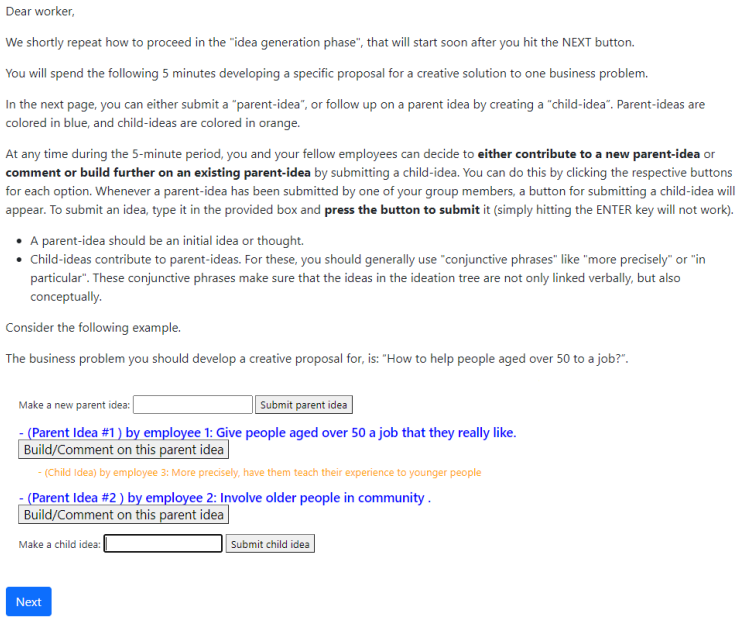
This study is a first attempt to examine peer evaluation characteristics and their effects on acceptance levels and team creativity. The results suggest that characteristics, and more specifically the purpose and format, of peer evaluations influence individuals' willingness to accept these systems, which might ultimately lead to behavioral shifts in creative performance. By providing this first exploratory piece of empirical evidence, we add to prior research on how peer evaluation characteristics such as format of peer evaluation presented to the manager (Arnold et al., 2018, 2020), and the purpose for which they are being used (Bamberger et al., 2005; Bettenhausen & Fedor, 1997; Fedor & Bettenhausen, 1989; Fedor et al., 1999; Hecht, Hobson, & Wang, 2020; Tavoletti et al., 2019) affect employees' willingness to accept these peer evaluations.

As with any experimental study, this study has its limitations. First, employee acceptability was measured in the post-experimental questionnaire. This might have led individuals to consider only the received peer evaluations from the last round when indicating the extent to which they agreed with the acceptance levels items. Future work could examine how acceptance levels change across periods. Second, we keep the social distance from the manager constant across our conditions. Nevertheless, prior research has shown that the social distance between an employee and manager affects workplace effort (Bohnet & Frey, 1999; Chen & Li, 2009). As such, future research might consider the

effect of peer evaluation purpose and/or format on employee outcomes under socially close and socially distant managers. This would especially be interesting given the current evolutions toward hybrid workplaces, where managers and employees might become more distant from each other. Third, while we focus on a specific situation where peer evaluations are used for non-monetary developmental purposes, we do acknowledge that the use of feedback messages by the managers in our study might be too artificial. The manner in which feedback is given from a manager to an employee is shown to affect the extent to which employees perceive this feedback as credible (Son & Kim, 2016). Moreover, prior research has shown that employees prefer personalized rewards from their managers, because they feel appreciated whenever a manager takes his/her time to acknowledge their employees (Bradler & Neckermann, 2019). Therefore, future work could examine how different ways of communicating final evaluations for developmental purposes (such as written, spoken, word use, and face-to-face) affect employee outcomes. Fourth, individuals in this study face an individual incentive scheme rather than a group incentive scheme. Prior research shows that group-based incentives promote cooperation among team members (Towry, 2003) which could eventually increase team creativity (Chen et al., 2012). Hence, future research might examine how an alternative incentive scheme such as group-based pay, might interact with peer evaluation purpose and affect group outcomes.

Appendices

Appendix 3.1: Screenshot of the experimental instructions



Dear worker,

We shortly repeat how to proceed in the "idea generation phase", that will start soon after you hit the NEXT button.

You will spend the following 5 minutes developing a specific proposal for a creative solution to one business problem.

In the next page, you can either submit a "parent-idea", or follow up on a parent idea by creating a "child-idea". Parent-ideas are colored in blue, and child-ideas are colored in orange.

At any time during the 5-minute period, you and your fellow employees can decide to **either contribute to a new parent-idea** or **comment or build further on an existing parent-idea** by submitting a child-idea. You can do this by clicking the respective buttons for each option. Whenever a parent-idea has been submitted by one of your group members, a button for submitting a child-idea will appear. To submit an idea, type it in the provided box and **press the button to submit** it (simply hitting the ENTER key will not work).

- A parent-idea should be an initial idea or thought.
- Child-ideas contribute to parent-ideas. For these, you should generally use "conjunctive phrases" like "more precisely" or "in particular". These conjunctive phrases make sure that the ideas in the ideation tree are not only linked verbally, but also conceptually.

Consider the following example.

The business problem you should develop a creative proposal for, is: "How to help people aged over 50 to a job?".

Make a new parent idea: Submit parent idea

- (Parent Idea #1) by employee 1: Give people aged over 50 a job that they really like.
Build/Comment on this parent idea

- (Child Idea) by employee 3: More precisely, have them teach their experience to younger people

- (Parent Idea #2) by employee 2: Involve older people in community .
Build/Comment on this parent idea

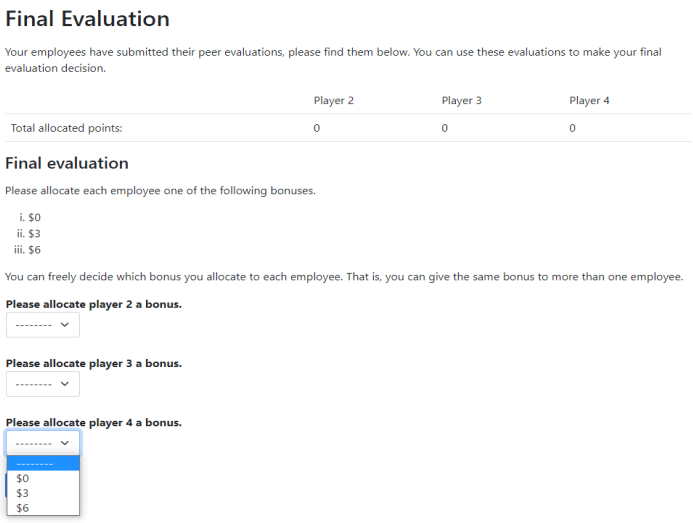
Make a child idea: Submit child idea

Next

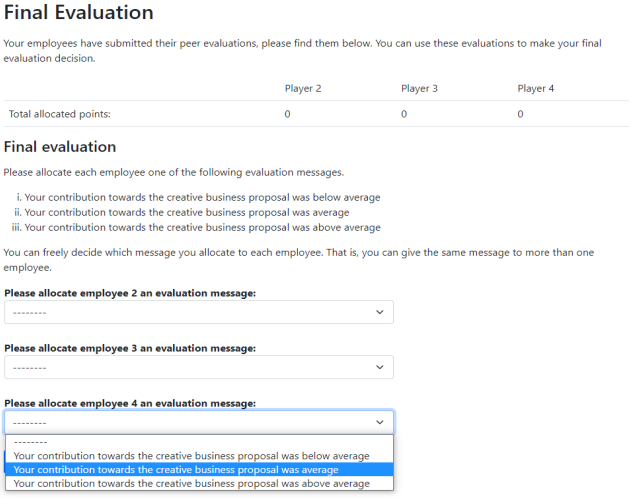
Source: author's screenshot including business problem adapted from Cardinaels et al., 2020.

Appendix 3.2: Manager task

Panel A: Final evaluation by the manager in the monetary purpose condition



Panel B: Final evaluation by the manager in the non-monetary purpose condition



Source: author's screenshot

Appendix 3.3: Coding scheme negotiation behavior

Category	Sub-code	Example
Integrative Tactics	1. Personal information exchange	1. A statement in which a group members reveals personal information like their name, place of residence.
	2. Rewards	2. A statement in which a group member creates a pleasant atmosphere (e.g. good work folks, thank you, fun working with you guys, I like the idea, lots of good ideas in this round).
	3. Request for cooperation	3. A statement in which a group members asks their group to cooperate (e.g. what do you think?, which idea?, which idea do we like?, I am open to any of the suggestions).
	4. Consensus	4. A statement in which a group member reach a consensus (e.g. sounds good, let's do this, me too, great idea).
Distributive Tactics	1. Punishment	1. A statement in which a group member creates an unpleasant atmosphere (e.g. I'm literally the only person doing the work, that's not helpful at all, it would actually help us get more of bonus if we actually discuss the topic, great I guess I am doing most of the work, we all need to agree, we need to make a decision).

Sub-code	Example
2. Persuasive arguments	2. A statement in which a group member aims to convince the their group to propose the same idea (e.g. I think we should go with 1, it is the most developed).
3. Refuse to share information	3. A statement in which a group member refuses to respond to a request for information (e.g. get lost).
4. Command	4. A statement in which a group member orders their group to put a proposal forward (e.g. okay then select 6, everyone pick number 8).

General Conclusion

The aim of this PhD dissertation was to improve our understanding of incentive program design on employee behavior. First, I discuss how the three chapters presented in this study contribute to the existing literature. Second, I describe the implications for business and practice of the findings from the three chapters. Finally, I point to the limitations of this dissertation as well as opportunities for future research.

Contribution to the literature

This dissertation makes several contributions to the literature. While incentive program design innovations such as offering employees reward choices and implementing peer evaluation systems occur frequently in practice, such design innovations have received little research attention in the domain of management control. Recently, this domain has gained some research interest, as shown by some recent influential publications (Arnold et al., 2018, 2019; Heninger et al., 2019; Holderness et al., 2017). We add to this upcoming research stream with three experimental studies.

The first study examines whether giving employees the opportunity to choose their reward, can affect their cognitive task performance. With this study, we contribute to the literature in several ways. First, we extend the scarce literature on the effects of reward choices given to individuals (Bareket-Bojmel et al., 2017; Caza et al., 20117; Williams & Luthans, 1992) by controlling for mental account in the reward choice and by considering different task types. Although prior research documents beneficial performance effects of administering individuals a reward choice, we expand this evidence by showing that reward choices only increase performance when they contain an extensive number of options. Second, while Bonner et al. (2000) argue that financial incentives do not motivate greater cognitive performance, we examine the effect of an alternative incentive on cognitive task performance. We show that performance on cognitively demanding tasks can be incentivized by offering individuals a reward choice containing tangible rewards.

The second study explores whether research investigating the effect of different types of managerial performance evaluations on employee effort can be generalized to peer performance evaluations. Indeed, our

findings suggest that when its outcomes are kept private, the use of rankings (also referred to as forced distribution ratings) incentivizes effort to a larger extent compared to the use of ratings in peer evaluations. However, we also show that the central tenet of prior literature showing that individuals work harder under forced distribution ratings rather than free ratings, does not generalize to settings where performance evaluation outcomes become transparent. Finally, we also contribute to the literature on peer monitoring (Falk & Ichino, 2006; Mas & Moretti, 2009; Towry, 2003) by demonstrating that the design of peer evaluations matters for their effectiveness as monitoring tools.

The third study investigates the role of narrative feedback in peer evaluations. While prior literature shows that the use of narrative feedback is beneficial for employee performance (Arnold et al., 2018, 2019, Lampe et al., 2021; Stubbs, 2021), our findings complement these studies by suggesting that the positive effect of narrative feedback depends on the purpose for which they are being used. Our findings suggest that the inclusion of narrative comments can be beneficial for employee acceptability levels when managers subsequently use them to determine bonus allocations. Nevertheless, our results also demonstrate that there is no relation between employee acceptability and creative performance, as we find that the use of narrative comments in monetary purpose peer evaluations deteriorates team creativity over time. Finally, with this study, we extend the literature on the use of management controls to incentivize creativity. Several accounting scholars have studied the effects of various individual incentives on creativity (Brüggen, Feichter, & Williamson, 2018; Cardinaels et al., 2020; Chen, Williamson, & Zhou, 2012; Kachelmeier, Bernhard, & Williamson, 2008). Nevertheless, the current literature remains scarce with respect to the effects of managerial evaluation systems on creativity, with Cardinaels & Feichter (2021) as a notable exception. We add to this stream of literature by showing that the design of peer performance evaluation systems can significantly affect team creative performance. We also extend this research stream to team creativity (Chen et al., 2012) instead of prior accounting research that predominantly focuses on the role of management control systems that affect individual creativity levels (Cardinaels, & Feichter, 2021; Cardinaels et al., 2021, Kachelmeier et al., 2008).

In summary, this dissertation shows that incentive program design affects a number of employee outcomes. For example, including reward choices in incentive programs can increase cognitive employee performance, but only when the reward choice covers a large number of options. In addition, when individuals are being evaluated by peers, they may be incentivized to work harder, but only in certain cases. Hence, it seems that peer evaluation design can affect employee effort positively, but under certain boundary conditions. Finally, in line with earlier research we find that the use of narrative comments in peer evaluations can positively affect performance outcomes, but we also show that it can have unintended consequences whenever managers tie bonuses to the outcomes of such evaluations.

Implications

The findings from this dissertation also provide important implications for management practice. For example, the results of the first study contribute to a better understanding of how a reward choice can affect employee performance. In particular, offering a reward choice to employees appears to increase their performance both on simple and more demanding tasks. However, we must note that our results might not generalize to all job types, as our results speak to rather boring, routine or low-skilled job types. Hence, our results are informative to managers leading employees in low-skilled jobs. An important implication of our study is that highly ambitious individuals appear to benefit more from being offered a reward choice. This suggests that managers should be aware of the possibility that offering reward choices in more demanding types of jobs, can have unintended consequences as highly ambitious individuals could grow into more senior or higher-skilled jobs.

In the second study, we gain insights about how the design of peer monitoring systems can affect employee effort in self-managing work groups. Our results indicate that, in the absence of a controlling manager, employees work harder when they evaluate their peers using a forced distribution rating system (i.e. ranking) when their organization uses a closed information policy. However, as organizations shift more to transparent settings (Bamberger & Belogolovsky, 2017), self-managing work groups might benefit more from rating each other rather than ranking each other. Specifically, the findings of this study demonstrate

that as transparency becomes more salient in the organization, employees have honesty concerns. Hence, organizations making use of such self-managing work groups might invest in an organizational culture that values honesty when it adopts an open information policy (Trevino, 1986).

With regard to the third study, the results are informative to organizations valuing creativity. We show that the inclusion of narrative comments in peer evaluations positively affects the extent to which employees accept the peer evaluation system and their group dynamics when managers use the outcomes of such evaluations to determine bonuses. Nevertheless, when an organization seeks to optimize creative performance, it is best to not tie monetary bonuses to peer evaluation outcomes but rather use those for more developmental purposes. As such, we argue that managers must be aware of the dangers involved in tying bonuses to evaluation outcomes, especially when they include narrative feedback as this may lead to high levels of impression management causing employees to lose track of what is really important to the organization.

Limitations and opportunities for future research

While all three studies provide contributions to the literature, we recognize that this dissertation is not without limitations. In what follows, I elaborate on these limitations by providing some interesting avenues for future research.

In the first study, we examine the effects of a reward choice in a gift-exchange setting without explicitly tying the reward to performance outcomes. Hence, future research can examine how different incentive structures affect the efficiency of a reward choice. Moreover, the job context might be more complex in reality. While we study the effects of a reward choice in a rather boring, low-skilled job setting, future research might study the effects in other types of job settings such as team-work settings or jobs that are characterized as intrinsically motivating.

Likewise, in our second study we employ one compensation type, namely a tournament compensation scheme. Future research could examine whether other types of compensation schemes might interact with peer evaluation design. Alternatively, in our second study we explicitly tie incentives towards both the quantity and quality dimensions

of employee effort. As such, future research could investigate how variations in both output dimensions are incentivized, and how this interacts with peer evaluation design. Finally, in this study we abstract away from the managerial role. Future studies could thus investigate whether our results hold for traditional work groups, where managers have a controlling role.

In the third study then, we recognize that our measure of employee acceptability needs further validation. That is, future research could study the effect of peer evaluation design on employee acceptability by executing a survey study. This way, our findings could be complemented by real practical evidence, and increase the generalizability of our results. Moreover, it would be interesting to examine whether and how the findings of this study would alter if individuals are compensated based on team outcomes rather than individual outcomes.

References

3D Group, (2013). Current Practices in 360 Degree Feedback: A Benchmark Study of North American Companies. Emeryville, CA.

Abeler, J., Falk, A., Goette, L., & Huffman, D. (2011). Reference Points and Effort Provision. The American economic review, 101(2), 470-492. doi:10.1257/aer.101.2.470

Adler, P. S., & Chen, C. X. (2011). Combining creativity and control: Understanding individual motivation in large-scale collaborative creativity. Accounting, organizations and society, 36(2), 63-85.

Ahn, T. S., Hwang, I., & Kim, M.-I. (2010). The Impact of Performance Measure Discriminability on Ratee Incentives. The Accounting review, 85(2), 389-417. doi:10.2308/accr.2010.85.2.389

Alicke, M. D., & Sedikides, C. (2009). Self-enhancement and self-protection: What they are and what they do. European review of social psychology, 20(1), 1-48. doi:10.1080/10463280802613866

Allen, B. J., Chandrasekaran, D., & Basuroy, S. (2018). Design Crowdsourcing: The Impact on New Product Performance of Sourcing Design Solutions from the "Crowd". Journal of marketing, 82(2), 106-123. https://doi.org/10.1509/jm.15.0481

Alonzo, V. (1996). The trouble with money. Incentive, 170, 26-33.

Amabile, T. M. (1996). Creativity in context: update to The social psychology of creativity: Boulder : Westview.

Anand, V., Webb, A., & Wong, C. (2018). Mitigating the Potentially Demotivating Effects of Early and Frequent Feedback About Goal Progress. Available at SSRN 3226304.

Anderson, S. W., Cheng, M. M., & Phua, Y. S. (2021). Influence of Control Precision and Prior Collaboration Experience on Trust and Cooperation in Inter-organizational Relationships. The Accounting Review. doi:10.2308/TAR-2019-0514

Andrew F. Hayes (Second edition. ed.): New York : Guilford Press.

Anseel, F. and F. Lievens (2006). Certainty as a moderator of feedback reactions? A test of the strength of the self-verification motive. Journal of Occupational and Organizational Psychology 79(4): 533-551.

Anseel, F., & Lievens, F. (2009). The mediating role of feedback acceptance in the relationship between feedback and attitudinal and performance outcomes. International Journal of Selection and Assessment, 17(4), 362-376.

Appelo, J. (2015). "The Peer-To-Peer Bonus System." Retrieved February 24th 2022, from https://www.forbes.com/sites/jurgenappelo/2015/07/08/the-peer-to-peer-bonus-system/?sh=1555932c4329.

Arber, M. M., Ireland, M. J., Feger, R., Marrington, J., Tehan, J., & Tehan, G. (2017). Ego Depletion in Real-Time: An Examination of the Sequential-Task Paradigm. Frontiers in Psychology, 8, 1672-1672. doi:10.3389/fpsyg.2017.01672

Ariely, D., Bracha, A., & Meier, S. (2009). Doing Good or Doing Well? Image Motivation and Monetary Incentives in Behaving Prosocially. American Economic Review, 99(1), 544-555. doi:10.1257/aer.99.1.544

Ariely, D., Kamenica, E., & Prelec, D. (2008). Man's search for meaning: The case of Legos. Journal of Economic Behavior & Organization, 67(3-4), 671-677. doi:10.1016/j.jebo.2008.01.004

Arnold, M. C., Hannan, L. R., & Tafkov, I. D. (2018). Team member subjective communication in homogeneous and heterogeneous teams. The Accounting Review, 93(5), 1-22. doi:10.2308/accr-52002

Arnold, M. C., Hannan, L. R., & Tafkov, I. D. (2020). Mutual Monitoring and Team Member Communication in Teams. The Accounting Review, 95(5), 1-21. doi:10.2308/accr-52659

Arnold, M. C., Hannan, R. L., & Tafkov, I. D. (2018). Team member subjective communication in homogeneous and heterogeneous teams. The Accounting review, 93(5), 1-22. doi:10.2308/accr-52002

Arnold, M. C., Hannan, R. L., & Tafkov, I. D. (2020). Mutual Monitoring and Team Member Communication in Teams. The Accounting Review, 95(5), 1-21. doi:10.2308/accr-52659

Arnold, M. C., Ponick, E., & Schenk-Mathes, H. Y. (2008). Groves mechanism vs. profit sharing for corporate budgeting–an experimental analysis with preplay communication. European Accounting Review, 17(1), 37-63.

Aronson, E. (1969). The theory of cognitive dissonance: A current perspective. In Advances in experimental social psychology (Vol. 4, pp. 1-34): Elsevier.

Baeten, X., & Verwaeren, B. (2012). Flexible Rewards From a Strategic Rewards Perspective. Compensation & Benefits Review, 44(1), 40-49. doi:10.1177/0886368712445541

Balafoutas, L., Czermak, S., Eulerich, M., & Fornwagner, H. (2020). Incentives for dishonisty: an experimental investigation with internal auditors. Economic inquiry, 58(2), 764-779. doi:10.1111/ecin.12878

Bamberger, P. A., Erev, I., Kimmel, M., & Oref-Chen, T. (2005). Peer assessment, individual performance, and contribution to group processes: The impact of rater anonymity. 30(4), 344-377. doi:10.1177/1059601104267619

Bamberger, P., & Belogolovsky, E. (2017). The dark side of transparency: How and when pay administration practices affect employee helping. Journal of Applied Psychology, 102(4), 658.

Bamberger, P., Erev, I., Kimmel, M., & Oref-Chen, T. (2005). Peer assessment, individual performance, and contribution to group processes: The impact of rater anonymity. Group & Organization Management, 30(4), 344-377. doi:10.1177/1059601104267619

Barber, A. E., Dunham, R. B., & Formisano, R. A. (1992). The Impact of Flexibe Benefits on Employee Motivation: a Field Study. Personnel Psychology, 45(1), 55-74. doi:10.1111/j.1744-6570.1992.tb00844.x

Bareket-Bojmel, L., Hochman, G., & Ariely, D. (2017). It's (Not) All About the Jacksons: Testing Different Types of Short-Term Bonuses in the Field. Journal of Management, 43(2), 534-554. doi:10.1177/0149206314535441

Barringer, M., & Milkovich, G. (1998). A theoretical exploration of the adoption and design of flexible benefit plans: A case of human resource innovation. Acad. Manage. Rev., 23(2), 305-324.

Baumeister, R. F. (1984). Choking under pressure: Self-consciousness and paradoxical effects of incentives on skillful performance. Journal of personality and social psychology, 46(3), 610-620. doi:10.1037//0022-3514.46.3.610

Baumeister, R. F., & Vohs, K. D. (2007). Self-Regulation, Ego Depletion, and Motivation. Social and Personality Psychology Compass, 1(1), 115-128. doi:10.1111/j.1751-9004.2007.00001.x

Baumeister, R. F., Bratslavsky, E., Finkenauer, C., & Vohs, K. D. (2001). Bad Is Stronger Than Good. Review of general psychology, 5(4), 323-370. doi:10.1037/1089-2680.5.4.323

Baumeister, R. F., Bratslavsky, E., Muraven, M., & Tice, D. M. (1998). Ego Depletion: Is the Active Self a Limited Resource? Journal of Personality and Social Psychology, 74(5), 1252-1265. doi:10.1037/0022-3514.74.5.1252

Baumeister, R. F., Smart, L., & Boden, J. M. (1996). Relation of threatened egotism to violence and aggression: The dark side of high self-esteem. . Psychological review, 103(1), 5-33. doi:10.1037/0033-295X.103.1.5

Bayus, B. L. (2013). Crowdsourcing New Product Ideas over Time: An Analysis of the Dell IdeaStorm Community. Management Science, 59(1), 226-244. https://doi.org/10.1287/mnsc.1120.1599

Becker, F., & Weißenberger, B. E. (2021). A Boo is Louder Than a Cheer: How Rejection and Feedback Type Influence Misreporting. Available at SSRN 3809631.

Becker, S. O., Messer, D., & Wolter, S. C. (2013). A Gift is Not Always a Gift: Heterogeneity and Long-term Effects in a Gift Exchange Experiment. Economica (London), 80(318), 345-371. doi:10.1111/ecca.12004

Beckers, N., Cardinaels, E., Dierynck, B., & Yin, H. (2018). How managers' on the job experience affects compensation design. Avaialble at SSRN: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3233645.

Bell, S. T., & Arthur Jr, W. (2008). Feedback acceptance in developmental assessment centers: The role of feedback message, participant personality, and affective response to the feedback session. Journal of Organizational Behavior: The International Journal of Industrial, Occupational and Organizational Psychology and Behavior, 29(5), 681-703.

Benistant, J., & Villeval, M. C. (2019). Unethical behavior and group identity in contests. Journal of economic psychology, 72, 128-155. doi:10.1016/j.joep.2019.03.001

Bentley, J. W. (2019). Decreasing operational distortion and surrogation through narrative reporting. The Accounting Review, 94(3), 27-55.
Bentley, J. W. (2021). Improving the Statistical Power and Reliability of Research Using Amazon Mechanical Turk. Accounting horizons. doi:10.2308/HORIZONS-18-052
Berg, J., Dickhaut, J., & McCabe, K. (1995). Trust, Reciprocity, and Social History. Games and Economic Behavior, 10(1), 122-142. doi:10.1006/game.1995.1027
Berger, J., Harbring, C., & Sliwka, D. (2013). Performance Appraisals and the Impact of Forced Distribution—An Experimental Investigation. Management Science, 59(1), 54-68. doi:10.1287/mnsc.1120.1624
Bettenhausen, K., & Fedor, D. (1997). Peer and upward appraisals: A comparison of their benefits and problems. Group & Organization Management, 22(2), 236-263.
Bicchieri, C. (2005). The Grammar of Society: The Nature and Dynamics of Social Norms: Cambridge: Cambridge University Press.
Bloomfield, R., Nelson, M. W., & Soltes, E. (2016). Gathering data for archival, field, survey, and experimental accounting research. Journal of Accounting Research, 54(2), 341-395.
Bohl, D. L. (1996). Minisurvey: 360-Degree Appraisals Yield Supervisor Results, Survey Shows,. Compensation and Benefits Review, 28(5), 16-19. doi:10.1177/088636879602800502
Bohnet, I., & Frey, B. S. (1999). Social distance and other-regarding behavior in dictator games: Comment. American Economic Review, 89(1), 335-339.
Bol, J. C. (2008). Subjectivity in compensation contracting. Paper presented at the AAA Management Accounting Section (MAS) 2006 Meeting Paper, Journal of Accounting Literature.
Bol, J. C. (2011). The determinants and performance effects of managers' performance evaluation biases. The Accounting Review, 86(5), 1549-1575.
Bol, J. C. (2011). The Determinants and Performance Effects of Managers' Performance Evaluation Biases. The Accounting review, 86(5), 1549-1575. doi:10.2308/accr-10099
Bol, J. C., Kramer, S., & Maas, V. S. (2016). How control system design affects performance evaluation compression: The role of

information accuracy and outcome transparency. Accounting, organizations and society, 51, 64-73. doi:10.1016/j.aos.2016.01.001

Bonner, S., & Sprinkle, G. (2002). The effects of monetary incentives on effort and task performance: theories, evidence, and a framework for research. Accounting, Organizations and Society, 27(4), 303-345. doi:10.1016/S0361-3682(01)00052-6

Bonner, S., Hastie, R., Sprinkle, G., & Young, S. (2000). A review of the effects of financial incentives on performance in laboratory tasks: Implications for management accounting. Journal of Management Accounting Research, 12, 19-64.

Boosey, L., & Goerg, S. (2020). The timing of discretionary bonuses – effort, signals, and reciprocity. Games and Economic Behavior, 124, 254-280. doi:10.1016/j.geb.2020.08.010

Botti, S., & Iyengar, S. S. (2004). The Psychological Pleasure and Pain of Choosing: When People Prefer Choosing at the Cost of Subsequent Outcome Satisfaction. Journal of Personality and Social Psychology, 87(3), 312-326. doi:10.1037/0022-3514.87.3.312

Bradler, C., & Neckermann, S. (2019). The Magic of the Personal Touch: Field Experimental Evidence on Money and Appreciation as Gifts. The Scandinavian journal of economics, 121(3), 1189-1221. doi:10.1111/sjoe.12310

Bradler, C., Dur, R., Neckermann, S., & Non, A. (2016). Employee Recognition and Performance: A Field Experiment. Management Science, 62(11), 3085-3099. doi:10.1287/mnsc.2015.2291

Brehm, J., & Festinger, L. (1957). Pressures Toward Uniformity of Performance in Groups. Human relations (New York), 10(1), 85-91. doi:10.1177/001872675701000106

Brett, J. F., & Atwater, L. E. (2001). 360° feedback: Accuracy, reactions, and perceptions of usefulness. Journal of Applied Psychology, 86(5), 930.

Brock, J. M., Lange, A., & Leonard, K. L. (2018). Giving and promising gifts: Experimental evidence on reciprocity from the field. J Health Econ, 58, 188-201. doi:10.1016/j.jhealeco.2018.02.007

Brüggen, A., & Strobel, M. (2007). Real effort versus chosen effort in experiments. Economics Letters, 96(2), 232-236.

Brüggen, A., Feichter, C., & Williamson, M. (2018). The Effect of Input and Output Targets for Routine Tasks on Creative Task Performance. The Accounting Review, 93(1), 29-43. doi:10.2308/accr-51781

Brun, J.-P., and Dugas, N. (2002), La reconnaissance au travail: une pratique riche de sens, Quebec, Canada: Chair in Occupational Health and Safety Management.

Brutus, S. (2010). Words versus numbers: A theoretical exploration of giving and receiving narrative comments in performance appraisal. Human Resource Management Review, 20(2), 144-157. doi:10.1016/j.hrmr.2009.06.003

Brutus, S., & Donia, M. M. (2010). Improving the effectiveness of students in groups with a centralized peer evaluation system. 9(4), 652-662. doi:10.5465/AMLE.2010.56659882

Buckless, F. A., & Ravenscroft, S. P. (1990). Contrast Coding: A Refinement of ANOVA in Behavioral Analysis. The Accounting Review, 65(4), 933-945. Retrieved from https://go.exlibris.link/76GvSYXW

Buunk, B. P., Zurriaga, R., Peiro, J. M., Nauta, A., & Gosalvez, I. (2005). Social comparisons at work as related to a cooperative social climate and to individual differences in social comparison orientation. Applied Psychology, 54(1), 61-80. doi:10.1111/j.1464-0597.2005.00196.x

Campbell, W. K., Reeder, G. D., Sedikides, C., & Elliot, A. J. (2000). Narcissism and Comparative Self-Enhancement Strategies. Journal of Research in Personality, 34(3), 329-347. doi:10.1006/jrpe.2000.2282

Cappelli, P., & Tavis, A. (2016). The performance management revolution. Harvard Business Review, 94(10), 58-67.

Cardinaels, E., & Feichter, C. (2021). Forced Rating Systems From Employee and Supervisor Perspectives. Journal of accounting research. doi:10.1111/1475-679X.12388

Cardinaels, E., Chen, C. X., & Yin, H. (2018). Leveling the playing field: The selection and motivation effects of tournament prize spread information. The Accounting review, 93(4), 127-149. doi:10.2308/accr-51955

Cardinaels, E., Dierynck, B., & Hu, W. (2020). Rejections, Incentives and Employee Creativity: When Chocolate Is Better Than Cash. [Working paper].

Carnevale, J. B., Huang, L., Vincent, L. C., Farmer, S., & Wang, L. (2021). Better to give than to receive (or seek) help? The interpersonal dynamics of maintaining a reputation for creativity. Organizational Behavior and Human Decision Processes, 167, 144-156.

Carpenter, J., Matthews, & Schirm. (2010). Tournaments and Office Politics: Evidence from a Real Effort Experiment. American Economic Review, 100(1), 504-517. doi:10.1257/aer.100.1.504

Carson, M. (2006). Saying it like it isn't: The pros and cons of 360-degree feedback. Business horizons, 49(5), 395-402. doi:10.1016/j.bushor.2006.01.004

Caza, A., McCarter, M. W., & Northcraft, G. B. (2015). Performance benefits of reward choice: a procedural justice perspective. Human Resource Management Journal, 25(2), 184-199. doi:10.1111/1748-8583.12073

Caza, A., McCarter, M. W., & Northcraft, G. B. (2015). Performance benefits of reward choice: A procedural justice perspective. Human Resource Management Journal, 25(2), 184-199.

Chan, E. W. (2018). Promotion, relative performance information, and the peter principle. The Accounting Review, 93(3), 83-103. doi:10.2308/accr-51890

Chang, L. J., Cheng, M. M., & Trotman, K. T. (2013). The effect of outcome and process accountability on customer–supplier negotiations. Accounting, Organizations and Society, 38(2), 93-107. doi:10.1016/j.aos.2012.12.002

Charness, G. (2000). Responsibility and effort in an experimental labor market. Journal of Economic Behavior and Organization, 42(3), 375-384. doi:10.1016/S0167-2681(00)00096-2

Charness, G., & Grieco, D. (2019). Creativity and incentives. Journal of the European Economic Association, 17(2), 454-496.

Charness, G., Masclet, D., & Villeval, M. C. (2014). The Dark Side of Competition for Status. Management Science, 60(1), 38-55. doi:10.1287/mnsc.2013.1747

Chen, C. X., & Sandino, T. (2012). Can Wages Buy Honesty? The Relationship Between Relative Wages and Employee Theft. Journal of accounting research, 50(4), 967-1000. doi:10.1111/j.1475-679X.2012.00456.x

Chen, C. X., Williamson, M. G., & Zhou, F. H. (2012). Reward System Design and Group Creativity: An Experimental Investigation. The Accounting review, 87(6), 1885-1911. doi:10.2308/accr-50232

Chen, D. L., Schonger, M., & Wickens, C. (2016). oTree—An open-source platform for laboratory, online, and field experiments. Journal of behavioral and experimental finance, 9, 88-97. doi:10.1016/j.jbef.2015.12.001

Chen, Y., & Li, S. X. (2009). Group identity and social preferences. American Economic Review, 99(1), 431-457.

Cheng, M. M., & Hsieh, C. (2009). Transfer Price Negotiation in the Presence of Unequal Bargaining Power: The Effect of a Peer Evaluation Scheme on Inter-divisional Profit Distribution. Australian Accounting Review, 19(3), 195-206. doi:10.1111/j.1835-2561.2009.00057.x

Chi, N.-W., & Lam, L. W. (2021). Is Negative Group Affective Tone Always Bad For Team Creativity? Team Trait Learning Goal Orientation as the Boundary Condition. Group & Organization Management, 10596011211011336.

Chiang, F., & Birtch, T. (2006). An empirical examination of reward preferences within and across national settings. Management International Review, 46(5), 573-596. doi:10.1007/s11575-006-0116-4

Choi, J. W., & Presslee, A. (2022). When and why tangible rewards can motivate greater effort than cash rewards: An analysis of four attribute differences. Accounting, Organizations and Society, 101389.

Choi, J. W., Hecht, G. W., & Tayler, W. B. (2012). Lost in Translation: The Effects of Incentive Compensation on Strategy Surrogation. The Accounting Review, 87(4), 1135-1163. doi:10.2308/accr-10273

Choi, J. W., Hecht, G. W., & Tayler, W. B. (2013). Strategy Selection, Surrogation, and Strategic Performance Measurement Systems: strategy selection and surrogation. Journal of Accounting

Research, 51(1), 105-133. doi:10.1111/j.1475-679X.2012.00465.x
Choi, W., Clark, J., & Presslee, A. (2019). Testing the effect of incentives on effort intensity using real-effort tasks.
Chow, C. W. (1983). 1983 Competitive Manuscript Award: The Effects of Job Standard Tightness and Compensation Scheme on Performance: An Exploration of Linkages. The Accounting Review, 58(4), 667-685. doi:10.2307/247062
Christ, M. H., Sedatole, K. L., & Towry, K. L. (2012). Sticks and carrots: The effect of contract frame on effort in incomplete contracts. The Accounting Review, 87(6), 1913-1938.
Christ, M. H., Sedatole, K. L., Towry, K. L., & Thomas, M. A. (2008). When formal controls undermine trust and cooperation. Strategic finance, 89(7), 39.
Chua, R. Y.-J., & Iyengar, S. S. (2006). Empowerment through Choice? A Critical Analysis of the Effects of Choice in Organizations. Research in Organizational Behavior, 27(C), 41-79. doi:10.1016/S0191-3085(06)27002-3
Church, B. K., Libby, T., & Zhang, P. (2008). Contracting frame and individual behavior: Experimental evidence. Journal of Management Accounting Research, 20(1), 153-168.
Clor-Proell, S. M., Guggenmos, R. D., & Rennekamp, K. (2020). Mobile Devices and Investment News Apps: The Effects of Information Release, Push Notification, and the Fear of Missing Out. The Accounting Review, 95(5), 95-115. doi:10.2308/accr-52625
Cole, M. S., Walter, F., & Bruch, H. (2008). Affective mechanisms linking dysfunctional behavior to performance in work teams: a moderated mediation study. Journal of Applied Psychology, 93(5), 945.
Cole, N. D., & Flint, D. H. (2004). Perceptions of distributive and procedural justice in employee benefits: flexible versus traditional benefit plans. Journal of Managerial Psychology, 19(1), 19-40. doi:10.1108/02683940410520646
Colella, A., Paetzold, R. L., Zardkoohi, A., & Wesson, M. J. (2007). Exposing Pay Secrecy. The Academy of Management review, 32(1), 55-71. doi:10.5465/AMR.2007.23463701
Coletti, A. L., Sedatole, K. L., & Towry, K. L. (2005). The Effect of Control Systems on Trust and Cooperation in Collaborative

Environments. The Accounting Review, 80(2), 477-500. doi:10.2308/accr.2005.80.2.477
Colquitt, J. A. (2012). Organizational justice.
Conway, J. M., & Huffcutt, A. I. (1997). Psychometric properties of multisource performance ratings: A meta-analysis of subordinate, supervisor, peer, and self-ratings. Human Performance, 10(4), 331-360.
Cook, T. D., Campbell, D. T., & Shadish, W. (2002). Experimental and quasi-experimental designs for generalized causal inference. Houghton Mifflin Boston, MA.
Cox, A. (2000). The importance of employee participation in determining pay system effectiveness. International Journal of Management Reviews, 2(4), 357-375.
David, E. M. (2013). Examining the Role of Narrative Performance Appraisal Comments on Performance. Human Performance, 26(5), 430-450. doi:10.1080/08959285.2013.836197
Davis, D., & Leo, R. A. (2012). INTERROGATION-RELATED REGULATORY DECLINE: Ego Depletion, Failures of Self-Regulation, and the Decision to Confess. Psychology, Public Policy, and Law, 18(4), 673-704. doi:10.1037/a0027367
De Dreu, C. K., & McCusker, C. (1997). Gain–loss frames and cooperation in two-person social dilemmas: A transformational analysis. Journal of Personality and Social Psychology, 72(5), 1093.
De Dreu, C. K., Baas, M., & Nijstad, B. A. (2008). Hedonic tone and activation level in the mood-creativity link: toward a dual pathway to creativity model. Journal of Personality and Social Psychology, 94(5), 739.
de la Torre-Ruiz, J. M., Vidal-Salazar, M. D., & Cordón-Pozo, E. (2019). Employees are satisfied with their benefits, but so what? The consequences of benefit satisfaction on employees' organizational commitment and turnover intentions. The International Journal of Human Resource Management, 30(13), 2097-2120.
De Stobbeleir, K., Ashford, S., & Buyens, D. (2011). Self-regulation of creativity at work: the role of feedback-seeking behavior in creative performance. Academy of Management Journal 54(4): 811-831.

De Stobbeleir, K., Ashford, S., & Zhang, C. (2020). Shifting focus: Antecedents and outcomes of proactive feedback seeking from peers. Human relations (New York), 73(3), 303-325. doi:10.1177/0018726719828448

De Vries, G. J., Gentile, E., Miroudot, S., & Wacker, K. M. (2020). The rise of robots and the fall of routine jobs. Labour Economics, 66, 101885.

Deci, E. L. (1971). Effects of externally mediated rewards on intrinsic motivation. Journal of personality and Social Psychology, 18(1), 105.

Deci, E. L., & Ryan, R. M. (1985). Intrinsic motivation and self-determination in human behavior: New York (N.Y.) : Plenum.

Deci, E. L., Koestner, R., & Ryan, R. M. (1999). A meta-analytic review of experiments examining the effects of extrinsic rewards on intrinsic motivation. Psychological bulletin, 125(6), 627.

DeNisi, A. S., & Kluger, A. N. (2000). Feedback effectiveness: Can 360-degree appraisals be improved? The Academy of Management executive (1993), 14(1), 129-139. doi:10.5465/AME.2000.2909845

DeNisi, A. S., Randolph, A. W., & Blencoe, A. G. (1983). Potential Problems with Peer Ratings. 26(3), 457-464. doi:10.2307/256256

Dewaele, L., Cardinaels, E., & Van den Abbeele, A. (2022). Peer evaluations: design characteristics and the impact on employee effort. Working paper. KU Leuven.

Dreisbach, G. (2012). Mechanisms of cognitive control: The functional role of task rules. Current Directions in Psychological Science, 21(4), 227-231.

Drory, A., & Zaidman, N. (2007). Impression management behavior: effects of the organizational system. Journal of Managerial Psychology.

Druskat, V. U., & Wolff, S. B. (1999). Effects and Timing of Developmental Peer Appraisals in Self-Managing Work Groups. Journal of Applied Psychology, 84(1), 58-74. https://doi.org/10.1037/0021-9010.84.1.58

Dzuranin, A. C., Randolph, D., & Stuart, N. V. (2013). Doing More With Less: Using Noncash Incentives to Improve Employee Performance. Journal of Corporate Accounting & Finance, 24(5), 75-80. doi:10.1002/jcaf.21877

Edelman, B., & Larkin, I. (2015). Social Comparisons and Deception Across Workplace Hierarchies: Field and Experimental Evidence. Organization Science, 26(1), 78-98. doi:10.1287/orsc.2014.0938

Edwards, M. R., & Ewen, A. J. (1996). How to Manage Performance and Pay with 360-Degree Feedback. Compensation and Benefits Review, 28(3), 41-46. doi:10.1177/088636879602800308

Edwin A. Locke, Gary P. Latham ; contr. by Ken J. Smith, a. o: Englewood Cliffs (N.J.) : Prentice Hall.

Ellemers, N., De Gilder, D., & Haslam, S. A. (2004). Motivating individuals and groups at work: A social identity perspective on leadership and group performance. Academy of Management Review, 29(3), 459-478.

Erez, A., Lepine, J. A., & Elms, H. (2002). Effects of rotated leadership and peer evaluation on the functioning and effectiveness of self-managed teams: a quasi-experiment., 55(4), 929-948. doi:10.1111/j.1744-6570.2002.tb00135.x

Essa, S. A. G., Dekker, H. C., & Groot, T. L. C. M. (2018). Your gain my pain? The effects of accounting information in uncertain negotiations. Management accounting research, 41, 20-42. doi:10.1016/j.mar.2018.02.002

Etzel, J. A., Cole, M. W., Zacks, J. M., Kay, K. N., & Braver, T. S. (2016). Reward motivation enhances task coding in frontoparietal cortex. Cerebral cortex, 26(4), 1647-1659.

Evans, J. H., Moser, D. V., Newman, A. H., & Stikeleather, B. R. (2016). Honor Among Thieves: Open Internal Reporting and Managerial Collusion. Contemporary accounting research, 33(4), 1375-1402. doi:10.1111/1911-3846.12181

Faillo, M., Grieco, D., & Zarri, L. (2019). The impact of peer ratings on cooperation: The role of information and cost of rating. Journal of public economic theory, 22(2), 408-432. doi:10.1111/jpet.12384

Falk, A., & Heckman, J. J. (2009). Lab experiments are a major source of knowledge in the social sciences. science, 326(5952), 535-538.

Falk, A., & Ichino, A. (2006). Clean Evidence on Peer Effects. Journal of labor economics, 24(1), 39-57. doi:10.1086/497818

Farrell, A. M., Kadous, K., & Towry, K. L. (2008). Contracting on Contemporaneous versus Forward-Looking Measures: An

Experimental Investigation. Contemporary Accounting Research, 25(3), 773-802. doi:10.1506/car.25.3.5
Fay, C. H., & Thompson, M. A. (2001). Contextual Determinants of Reward Systems' Success: An Exploratory Study. Human Resource Management, 40(3), 213-226. doi:10.1002/hrm.1012
Fedor, D. B., & Bettenhausen, K. L. (1989). The Impact of Purpose, Participant Preconceptions, and Rating Level on the Acceptance of Peer Evaluations. Group & Organization Management, 14(2), 182-197. doi:10.1177/105960118901400207
Fedor, D. B., Bettenhausen, K. L., & Davis, W. (1999). Peer Reviews: Employees' Dual Roles as Raters and Recipients. In (Vol. 24, pp. 92-120).
Fehr, E., & Gachter, S. (2000). Fairness and Retaliation: The Economics of Reciprocity. The Journal of economic perspectives, 14(3), 159-181. doi:10.1257/jep.14.3.159
Ferreira, A., & Otley, D. (2009). The design and use of performance management systems: An extended framework for analysis. Management accounting research, 20(4), 263-282.
Fisher, J. G., Maines, L. A., Peffer, S. A., & Sprinkle, G. B. (2002). Using budgets for performance evaluation: Effects of resource allocation and horizontal information asymmetry on budget proposals, budget slack, and performance. The Accounting Review, 77(4), 847-865.
Futrell, C. M., & Jenkins, O. C. (1978). Pay Secrecy versus Pay Disclosure for Salesmen: A Longitudinal Study. Journal of marketing research, 15(2), 214-219. doi:10.1177/002224377801500204
Gailliot, M. T., Baumeister, R. F., Dewall, C. N., Maner, J. K., Plant, E. A., Tice, D. M., . . . Schmeichel, B. J. (2007). Self-Control Relies on Glucose as a Limited Energy Source: Willpower Is More Than a Metaphor. Journal of Personality and Social Psychology, 92(2), 325-336. doi:10.1037/0022-3514.92.2.325
Gallus, J. (2017). Fostering Public Good Contributions with Symbolic Awards: A Large-Scale Natural Field Experiment at Wikipedia. Management Science, 63(12), 3999-4015. doi:10.1287/mnsc.2016.2540
Gallus, J., Jung, O. S., & Lakhani, K. (2019). Managerial Recognition as an Incentive for Innovation Platform Engagement: A Field

Experiment and Interview Study at NASA. Harvard Business School Working Paper.
Gartner. (2018). Peer Feedback Boosts Performance. Retrieved from https://www.gartner.com/smarterwithgartner/peer-feedback-boosts-employee-performance/
Gibbs, M., Merchant, K. A., Stede, W. A. V. d., & Vargus, M. E. (2004). Determinants and effects of subjectivity in incentives. The Accounting Review, 79(2), 409-436.
Giebels, E., De Dreu, C. K. W., & Van De Vliert, E. (2000). Interdependence in negotiation: effects of exit options and social motive on distributive and integrative negotiation. European journal of social psychology, 30(2), 255-272. doi:10.1002/(SICI)1099-0992(200003/04)30:2<255::AID-EJSP991>3.0.CO;2-7
Gill, D., & Prowse, V. (2013). A Novel Computerized Real Effort Task Based on Sliders. IDEAS Working Paper Series from RePEc.
Givi, J., & Galak, J. (2017). Sentimental value and gift giving: Givers' fears of getting it wrong prevents them from getting it right. Journal of consumer psychology, 27(4), 473-479. doi:10.1016/j.jcps.2017.06.002
Glover, J. C., & Xue, H. (2020). Team Incentives and Bonus Floors in Relational Contracts. The Accounting Review, Forthcoming https://doi.org/10.2308/tar-2016-0630.
Gorman, C. A., Meriac, J. P., Roch, S. G., Ray, J. L., & Gamble, J. S. (2017). An exploratory study of current performance management practices: Human resource executives' perspectives. International Journal of Selection and Assessment, 25(2), 193-202. doi:10.1111/ijsa.12172
Grabner, I. (2014). Incentive system design in creativity-dependent firms. The Accounting Review, 89(5), 1729-1750.
Graham, J. L., Evenko, L. I., & Rajan, M. N. (1992). An empirical comparison of Soviet and American business negotiations. Journal of International Business Studies, 23(3), 387-418.
Groen, B. A. (2018). A survey study into participation in goal setting, fairness, and goal commitment: Effects of including multiple types of fairness. Journal of Management Accounting Research, 30(2), 207-240.

Guggenmos, R. D., Piercey, M. D., & Agoglia, C. P. (2018). Custom contrast testing: Current trends and a new approach. The Accounting Review, 93(5), 223-244.

Gürtler, O., & Harbring, C. (2010). Feedback in Tournaments under Commitment Problems: Experimental Evidence. Journal of economics & management strategy, 19(3), 771-810. doi:10.1111/j.1530-9134.2010.00269.x

Haden, J. (2017). 25 rewards that great employees actually love to receive. Retrieved from https://www.inc.com/jeff-haden/25-creative-rewards-that-great-employees-actually-love-to-receive.html

Haesebrouck, K., Cools, M., & Van den Abbeele, A. (2018). Status Differences and Knowledge Transfer: The Effect of Incentives. The Accounting review, 93(1), 213-234. doi:10.2308/accr-51765

Hajcak, G., Holroyd, C. B., Moser, J. S., & Simons, R. F. (2005). Brain potentials associated with expected and unexpected good and bad outcomes. Psychophysiology, 42(2), 161-170.

Hales, J., Wang, L., & Williamson, M. G. (2015). Selection Benefits of Stock-Based Compensation for the Rank-and-File. Account. Rev., 90(4), 1497-1516. doi:10.2308/accr-50962

Hall-McMaster, S., Muhle-Karbe, P. S., Myers, N. E., & Stokes, M. G. (2019). Reward boosts neural coding of task rules to optimize cognitive flexibility. Journal of Neuroscience, 39(43), 8549-8561.

Hammermann, A., & Mohnen, A. (2014). Who benefits from benefits? Empirical research on tangible incentives. Review of managerial science, 8(3), 327-350. doi:10.1007/s11846-013-0107-3

Hannan, R. L., Krishnan, R., & Newman, A. H. (2008). The Effects of Disseminating Relative Performance Feedback in Tournament and Individual Performance Compensation Plans. The Accounting Review, 83(4), 893-913. doi:10.2308/accr.2008.83.4.893

Hannan, R. L., McPhee, G. P., Newman, A. H., & Tafkov, I. D. (2013). The effect of relative performance information on performance and effort allocation in a multi-task environment.(Report). Accounting Review, 88(2), 553. doi:10.2308/accr-50312

Hannan, R. L., Towry, K. L., & Zhang, Y. (2013). Turning Up the Volume: An Experimental Investigation of the Role of Mutual

Monitoring in Tournaments. Contemporary Accounting Research, 30(4), 1401-1426. doi:10.1111/1911-3846.12006
Harbring, & Irlenbusch. (2011). Sabotage in Tournaments: Evidence from a Laboratory Experiment. Management Science, 57(4), 611-627. https://doi.org/10.1287/mnsc.1100.1296
Harris, S., Mussen, P., & Rutherford, E. (1976). Some cognitive, behavioral and personality correlates of maturity of moral judgment. The Journal of genetic psychology, 128(1st Half), 123-135. Retrieved from https://go.exlibris.link/DgBg3dPt
Hayes, A. F. (2018). Introduction to mediation, moderation, and conditional process analysis: a regression-based approach
Hecht, G., Hobson, J. L., & Wang, L. W. (2020). The effect of performance reporting frequency on employee performance. The Accounting Review, 95(4), 199-218. doi:10.2308/ACCR-52601
Heneman, R. L., Fisher, M. M., & Dixon, K. E. (2001). Reward and Organizational Systems Alignment: An Expert System. Compensation & Benefits Review, 33(6), 18-29. doi:10.1177/08863680122098694
Heninger, W. G., Smith, S. D., & Wood, D. A. (2019). Reward type and performance: An examination of organizational wellness programs. Management accounting research, 44, 1-11. doi:10.1016/j.mar.2019.02.001
Hesford, J., Mangin, N., & Pizzini, M. (2020). Using Fixed Wages for Management Control: An Intra-Firm Test of the Effect of Relative Compensation on Performance. Journal of Management Accounting Research, 32(3), 137-154. doi:10.2308/jmar-18-070
Higgins, E. T. (1996). The “Self Digest”: Self-Knowledge Serving Self-Regulatory Functions. Journal of Personality and Social Psychology, 71(6), 1062-1083. doi:10.1037/0022-3514.71.6.1062
Hillebrink, C., Schippers, J., van Doorne-Huiskes, A., & Peters, P. (2008). Offering choice in benefits: a new Dutch HRM arrangement. International Journal of Manpower, 29(4), 304-322. doi:10.1108/01437720810884737
Hodosh, M., Young, P., & Hockenmaier, J. (2013). Framing Image Description as a Ranking Task: Data, Models and Evaluation Metrics. The Journal of artificial intelligence research, 47, 853-899. doi:10.1613/jair.3994

Hoffman, E., McCabe, K., & Smith, V. L. (1996). Social Distance and Other-Regarding Behavior in Dictator Games. The American Economic Review, 86(3), 653-660.

Hofmann, D. A., Lei, Z., & Grant, A. M. (2009). Seeking help in the shadow of doubt: The sensemaking processes underlying how nurses decide whom to ask for advice. Journal of Applied Psychology, 94(5), 1261.

Holderness Jr, D. K., Olsen, K. J., & Thornock, T. A. (2017). Who are you to tell me that?! The moderating effect of performance feedback source and psychological entitlement on individual performance. Journal of Management Accounting Research, 29(2), 33-46.

Homem de Mello, F. (2019). Google's Performance Management Practices. Retrieved from https://qulture.rocks/en/blog/googles-performance-management-practices-part-1/

Höppe, F., & Moers, F. (2011). The choice of different types of subjectivity in CEO annual bonus contracts. The Accounting Review, 86(6), 2023-2046.

Huang, S.-W., & Fu, W.-T. (2013, 2013). Don't hide in the crowd: increasing social transparency between peer workers improves crowdsourcing outcomes.

Ilgen, D. R., & Davis, C. (2000). Bearing Bad News: Reactions to Negative Performance Feedback. Applied Psychology, 49(3), 550-565. doi:10.1111/1464-0597.00031

Ilgen, D. R., Fisher, C. D., & Taylor, M. S. (1979). Consequences of individual feedback on behavior in organizations. Journal of Applied Psychology, 64(4), 349-371. doi:10.1037/0021-9010.64.4.349

Incentive Federation Inc. (2016). Incentive Marketplace Estimate Research Study. Retrieved from http://www.incentivefederation.org/wp-content/uploads/2016/07/Incentive-Marketplace-Estimate-Research-Study-2015-16-White-Paper.pdf

Incentive Research Foundation. (2015). Landmark study: participant award experience preferences. Retrieved from https://theirf.org/research/2015-ima-landmark-study-participant-award-experience-preferences/1619/

Incentive Research Foundation. (2017). Conscious and Unconscious Reward Preference & Choice: A Biometric Experiment. Retrieved from http://theirf.org/research/conscious-and-unconscious-reward-preference-choice-a-biometric-experiment/2328/

Incentive Research Foundation. (2018). A Closer Look at Gift Cards: U.S. Spend, Support, Sourcing, and Services for Gift Card Programs in Corporate Organizations. Retrieved from https://theirf.org/research/a-closer-look-at-gift-cards-us-spend-support-sourcing-and-services-for-gift-card-programs-in-corporate-organizations/2409/

Iyengar, S. S., & Lepper, M. R. (2000). When Choice is Demotivating: Can One Desire Too Much of a Good Thing? Journal of Personality and Social Psychology, 79(6), 995-1006. doi:10.1037/0022-3514.79.6.995

Jackson, D. J. R., Michaelides, G., Dewberry, C., Schwencke, B., & Toms, S. (2020). The Implications of Unconfounding Multisource Performance Ratings. J Appl Psychol, 105(3), 312-329. doi:10.1037/apl0000434

Jawahar, I. M., & Williams, C. R. (1997). We are all the children above average: the performance appraisal purpose effect. Personnel psychology, 50(4), 905-925. doi:10.1111/j.1744-6570.1997.tb01487.x

Jeffrey, S. A. (2009). Justifiability and the Motivational Power of Tangible Noncash Incentives. Human Performance, 22(2), 143-155. doi:10.1080/08959280902743659

Jeffrey, S. A., & Shaffer, V. (2007). The Motivational Properties of Tangible Incentives. Compensation & Benefits Review, 39(3), 44-50. doi:10.1177/0886368707302528

Joo, B.-K., Song, J. H., Lim, D. H., & Yoon, S. W. (2012). Team creativity: the effects of perceived learning culture, developmental feedback and team cohesion. International Journal of Training and Development, 16(2), 77-91. doi:10.1111/j.1468-2419.2011.00395.x

Jordan, P. J., Lawrence, S. A., & Troth, A. C. (2006). The impact of negative mood on team performance. Journal of management & organization, 12(2), 131-145.

Kachelmeier, S. J., & Williamson, M. G. (2010). Attracting Creativity: The Initial and Aggregate Effects of Contract Selection on Creativity-Weighted Productivity. The Accounting Review, 85(5), 1669-1691.

Kachelmeier, S. J., Reichert, B. E., & Williamson, M. G. (2008). Measuring and Motivating Quantity, Creativity, or Both. Journal of accounting research, 46(2), 341-373. doi:10.1111/j.1475-679X.2008.00277.x

Kachelmeier, S. J., Thornock, T. A., & Williamson, M. G. (2016). Communicated Values as Informal Controls: Promoting Quality While Undermining Productivity? Contemporary Accounting Research, 33(4), 1411-1434. doi:10.1111/1911-3846.12147

Kadous, K., & Zhou, Y. D. (2017). Maximizing the contribution of JDM-style experiments in accounting. In The Routledge companion to behavioural accounting research (pp. 175-192). Routledge.

Kahneman, D., & Tversky, A. (1979). Prospect Theory: An Analysis of Decision under Risk. Econometrica, 47(2), 263-291. doi:10.2307/1914185

Kampkötter, P., & Sliwka, D. (2011). Differentiation and performance: An empirical investigation on the incentive effects of bonus plans. IZA Discussion Paper No. 6070, Available at SSRN: https://ssrn.com/abstract=1955410 or http://dx.doi.org/10.2139/ssrn.1955410

Kane, J. S., & Lawler, E. E. (1978). Methods of peer assessment. Psychological bulletin, 85(3), 555-586. doi:10.1037/0033-2909.85.3.555

Keh, H. T., & Lee, Y. H. (2006). Do reward programs build loyalty for services?: The moderating effect of satisfaction on type and timing of rewards. Journal of Retailing, 82(2), 127-136. doi:10.1016/j.jretai.2006.02.004

Kelley, H. H., & Stahelski, A. J. (1970). Social interaction basis of cooperators' and competitors' beliefs about others. Journal of Personality and Social Psychology, 16(1), 66.

Kelliher, C., & Anderson, D. (2008). For better or for worse? An analysis of how flexible working practices influence employees' perceptions of job quality. The International Journal of Human Resource Management, 19(3), 419-431. doi:10.1080/09585190801895502

Kelly, K., Dinovitzer, R., Gunz, H., & Gunz, S. P. (2020). The Interaction of Perceived Subjectivity and Pay Transparency on Professional Judgment in a Profit Pool Setting: The Case of Large Law Firms. The Accounting Review, 95(5), 227-246. doi:10.2308/accr-52612

Kelly, K., Presslee, A., & Webb, R. (2017). The Effects of Tangible Rewards versus Cash Rewards in Consecutive Sales Tournaments: A Field Experiment. Account. Rev., 92(6), 165-185. doi:10.2308/accr-51709

Kluger, A. N., & DeNisi, A. (1996). The Effects of Feedback Interventions on Performance: A Historical Review, a Meta-Analysis, and a Preliminary Feedback Intervention Theory. Psychological Bulletin, 119(2), 254-284. doi:10.1037/0033-2909.119.2.254

Koch, A. K., & Nafziger, J. (2016). Gift exchange, control, and cyberloafing: A real-effort experiment. Journal of Economic Behavior & Organization, 131, 409-426. doi:10.1016/j.jebo.2016.09.008

Koo, R. C. (2011). Global added value of flexible benefits. Benefits quarterly, 27(4), 17-20.

Kool, W., & Botvinick, M. (2018). Mental labour. Nature human behaviour, 2(12), 899-908.

Kool, W., McGuire, J. T., Rosen, Z. B., & Botvinick, M. M. (2010). Decision Making and the Avoidance of Cognitive Demand. J Exp Psychol Gen, 139(4), 665-682. doi:10.1037/a0020198

Krippendorff, K. (2018). Content analysis: An introduction to its methodology: Sage publications.

Kube, S., Marchal, M. A., & Puppe, C. (2012). The Currency of Reciprocity: Gift Exchange in the Workplace. American Economic Review, 102(4), 1644-1662. doi:10.1257/aer.102.4.1644

Kuhnen, C. M., & Tymula, A. (2012). Feedback, Self-Esteem, and Performance in Organizations. 58(1), 94-113. doi:10.1287/mnsc.1110.1379

Lampe, Shäffer, & Schaupp. (2021). The Performance Effects of Narrative Feedback. WHU - Otto Beisheim School of Management.

Lawler, E. E. (1990). Strategic pay : aligning organizational strategies and pay systems: San Francisco (Calif.) : Jossey-Bass.
Lawler, E. E., & Hackman, J. R. (1969). Impact of employee participation in the development of pay incentive plans: A field experiment. Journal of Applied Psychology, 53(6), 467.
Lazear, E. P. (1989). Pay Equality and Industrial Politics. Journal of Political Economy, 97(3), 561-580. doi:10.1086/261616
Lazear, E. P., & Rosen, S. (1981). Rank-Order Tournaments as Optimum Labor Contracts. The Journal of political economy, 89(5), 841-864. doi:10.1086/261010
Leana, C. R., Locke, E. A., & Schweiger, D. M. (1990). Fact and Fiction in Analyzing Research on Participative Decision Making: A Critique of Cotton, Vollrath, Froggatt, Lengnick-Hall, and Jennings. The Academy of Management Review, 15(1), 137-146. doi:10.2307/258110
Leibbrandt, A., Wang, L. C., & Foo, C. (2018). Gender Quotas, Competitions, and Peer Review: Experimental Evidence on the Backlash Against Women. Management science, 64(8), 3501-3516. doi:10.1287/mnsc.2017.2772
Locke, E. A., & Latham, G. P. (1990). A theory of goal setting and task performance
Lockwood, P., & Kunda, Z. (1997). Superstars and me: Predicting the impact of role models on the self. Journal of Personality and Social Psychology, 73(1), 91-103. doi:10.1037/0022-3514.73.1.91
Loftus, S., & Tanlu, L. J. (2018). Because of "because": Examining the use of causal language in relative performance feedback. The Accounting review, 93(2), 277-297. doi:10.2308/accr-51830
López-Pérez, R., & Vorsatz, M. (2010). On approval and disapproval: Theory and experiments. Journal of Economic Psychology, 31(4), 527-541. doi:10.1016/j.joep.2010.03.016
Lourenço, S. M. (2020). Do self-reported motivators really motivate higher performance? Management accounting research., 47, 100676. doi:10.1016/j.mar.2019.100676
Maas, V. S., & Van Rinsum, M. (2013). How Control System Design Influences Performance Misreporting. Journal of accounting research, 51(5), 1159-1186. doi:10.1111/1475-679x.12025

Madjar, N. (2005). The Contributions of Different Groups of Individuals to Employees' Creativity. Advances in developing human resources, 7(2), 182-206. doi:10.1177/1523422305274525

Maley, J. F., Dabic, M., & Moeller, M. (2020). Employee performance management: charting the field from 1998 to 2018. International Journal of Manpower, 42(1), 131-149. doi:10.1108/IJM-10-2019-0483

Marx, L. M., & Squintani, F. (2009). Individual accountability in teams. Journal of Economic Behavior and Organization, 72(1), 260-273. doi:10.1016/j.jebo.2009.05.009

Mas, A., & Moretti, E. (2009). Peers at Work. The american economic review, 99(1), 112-145. doi:10.1257/aer.99.1.112

Mazar, N., Amir, O., & Ariely, D. (2008). The Dishonesty of Honest People: A Theory of Self-Concept Maintenance. Journal of Marketing Research, 45(6), 633-644. https://doi.org/10.1509/jmkr.45.6.633

Merchant, K. A., & Van der Stede, W. A. (2017). Management control systems: performance measurement, evaluation and incentives. 4th edition. Pearson education.

Michalowicz, M. (2013). 101 ways to reward employees (without giving them cash). Retrieved from https://www.americanexpress.com/us/small-business/openforum/articles/a-101-ways-to-reward-employees-without-giving-them-cash/

Mitchell, T., Presslee, A., Schulz, A. K. D., & Webb, A. (2021). Needs Versus Wants: The Mental Accounting and Effort Effects of Tangible Rewards. Journal of Management Accounting Research. doi:10.2308/JMAR-2019-505

Moers, F. (2005). Discretion and bias in performance evaluation: the impact of diversity and subjectivity. Accounting, organizations and society, 30(1), 67-80. doi:10.1016/j.aos.2003.11.001

Morgeson, F., Campion, M., & Maertz, C. (2001). Understanding Pay Satisfaction: The Limits of a Compensation System Implementation. Journal of Business and Psychology, 16(1), 133-149. doi:10.1023/A:1007848007459

Mosley, E. (2015). Creating an Effective Peer Review System. Harvard Business Review. Creating an Effective Peer Review System.

Retrieved from https://hbr.org/2015/08/creating-an-effective-peer-review-system

Myers, L., Downie, S., Taylor, G., Marrington, J., Tehan, G., & Ireland, M. J. (2018). Understanding Performance Decrements in a Letter-Canceling Task: Overcoming Habits or Inhibition of Reading.(Report). Frontiers in Psychology, 9(MAY). doi:10.3389/fpsyg.2018.00711

Nadler, A., Ellis, S., & Bar, I. (2003). To seek or not to seek: The relationship between help seeking and job performance evaluations as moderated by task-relevant expertise. Journal of Applied Social Psychology, 33(1), 91-109.

Newman, A. H., & Tafkov, I. D. (2014). Relative performance information in tournaments with different prize structures. Accounting, Organizations and Society, 39(5), 348-361.

Ng, K.-Y., Koh, C., Ang, S., Kennedy, J. C., & Chan, K.-Y. (2011). Rating Leniency and Halo in Multisource Feedback Ratings: Testing Cultural Assumptions of Power Distance and Individualism-Collectivism. Journal of Applied Psychology, 96(5), 1033-1044. doi:10.1037/a0023368

Norberg, P. A. (2017). Employee Incentive Programs: Recipient Behaviors in Points, Cash, and Gift Card Programs. Performance Improvement Quarterly, 29(4), 375-388. doi:10.1002/piq.21233

Oehlmann, M., Meyerhoff, J., Mariel, P., & Weller, P. (2017). Uncovering context-induced status quo effects in choice experiments. Journal of environmental economics and management, 81, 59-73. doi:10.1016/j.jeem.2016.09.002

Otto, A. R., & Vassena, E. (2021). It's all relative: Reward-induced cognitive control modulation depends on context. Journal of Experimental Psychology: General, 150(2), 306.

Parro, C., Dixon, M. L., & Christoff, K. (2018). The neural basis of motivational influences on cognitive control. Human brain mapping, 39(12), 5097-5111.

Patall, Cooper, & Robinson. (2008). The Effects of Choice on Intrinsic Motivation and Related Outcomes: A Meta-Analysis of Research Findings. Psychological Bulletin, 134(2), 270-300. doi:10.1037/0033-2909.134.2.270

Peltier, J., Schultz, D., & Block, M. (2005). Awards Selection Study Phase I: Preliminary Insights from Managers Paper presented at the The Forum for People Performance Management and Measurement, Northwestern University. http://www.enterpriseengagement.org/pdf/awards_selection_study_phase1.pdf

Pennebaker, J. W., Boyd, R. L., Jordan, K., & Blackburn, K. (2015). The development and psychometric properties of LIWC2015. Retrieved from

People Driven Performance. (2015). Manufacturers Must Recognize and Consistently Award Employee Loyalty. Retrieved from http://pdpsolutions.com/manufacturers-must-recognize-and-consistently-award-employee-loyalty

Perkins, S. J., & Jones, S. (2020). Reward Management: Alternatives, Consequences and Contexts.: Konan Page Publishers.

Perry-Smith, J. E., & Shalley, C. E. (2003). The social side of creativity: A static and dynamic social network perspective. The Academy of Management Review, 28(1), 89-106. doi:10.5465/AMR.2003.8925236

Peterson, R. A. (1994). A meta-analysis of Cronbach's coefficient alpha. Journal of Consumer Research, 21(2), 381-391.

Pfeiffer, T., & Velthuis, L. (2009). Incentive system design based on accrual accounting: A summary and analysis. Journal of Management Accounting Research, 21(1), 19-53.

Prendergast, C. (1999). The provision of incentives in firms. Journal of economic literature, 37(1), 7-63.

Prendergast, C. (1999). The provision of incentives in firms. Journal of economic literature, 37(1), 7-63.

Presslee, A., Vance, T. W., & Webb, R. A. (2013). The effects of reward type on employee goal setting, goal commitment, and performance. Accounting Review, 88(5), 1805-1831. doi:10.2308/accr-50480

Pruitt, D. G. (1981). Negotiation behavior. New York: Academic Press.

Pruitt, D. G., & Lewis, S. A. (1975). Development of integrative solutions in bilateral negotiation. Journal of Personality and Social Psychology, 31(4), 621.

Robitschek, C., Ashton, M. W., Spering, C. C., Geiger, N., Byers, D., Schotts, G. C., & Thoen, M. A. (2012). Development and

Psychometric Evaluation of the Personal Growth Initiative Scale-II. Journal of Counseling Psychology, 59(2), 274-287. doi:10.1037/a0027310

Roch, S. G., Sternburgh, A. M., & Caputo, P. M. (2007). Absolute vs Relative Performance Rating Formats: Implications for fairness and organizational justice. International journal of selection and assessment, 15(3), 302-316. doi:10.1111/j.1468-2389.2007.00390.x

Rosaz, J., & Villeval, M. C. (2012). Lies and biased evaluation: A real-effort experiment. Journal of economic behavior & organization, 84(2), 537-549. doi:10.1016/j.jebo.2012.09.002

Rosenthal, R., & Rosnow, R. L. (1985). Contrast analysis: Focused comparisons in the analysis of variance: CUP Archive.

Rubin, J., Samek, A., & Sheremeta, R. M. (2018). Loss aversion and the quantity–quality tradeoff. Experimental Economics, 21(2), 292-315.

Ryan, R. M., Mims, V., & Koestner, R. (1983). Relation of reward contingency and interpersonal context to intrinsic motivation: A review and test using cognitive evaluation theory. Journal of Personality and Social Psychology, 45(4), 736-750. doi:10.1037/0022-3514.45.4.736

Rynes, S. L., Gerhart, B., & Parks, L. (2005). PERSONNEL PSYCHOLOGY: Performance. Annu. Rev. Psychol, 56, 571-600.

Saavedra, R., & Kwun, S. K. (1993). Peer Evaluation in Self-Managing Work Groups. Journal of Applied Psychology, 78(3), 450-462. https://doi.org/10.1037/0021-9010.78.3.450

Schleicher, D. J., Bull, R. A., & Green, S. G. (2009). Rater Reactions to Forced Distribution Rating Systems. Journal of management, 35(4), 899-927. doi:10.1177/0149206307312514

Schmeichel, B. J., & Zell, A. (2007). Trait self-control predicts performance on behavioral tests of self-control. Journal of personality, 75(4), 743-755.

Sela, A., Berger, J., & Liu, W. (2009). Variety, Vice, and Virtue: How Assortment Size Influences Option Choice. Journal of Consumer Research, 35(6), 941-951. doi:10.1086/593692

Shaffer, V. A., & Arkes, H. R. (2009). Preference reversals in evaluations of cash versus non-cash incentives. Journal of Economic Psychology, 30(6), 859-872. doi:10.1016/j.joep.2009.08.001

SHRM. (2018). Using Recognition and Other Workplace Efforts to Engage Employees. Retrieved from

Sol, J. (2016). Peer Evaluation: Incentives and Coworker Relations. Journal of Economics and Management Strategy, 25(1), 56-76. doi:10.1111/jems.12134

Son, S., & Kim, D.-Y. (2016). The role of perceived feedback sources' learning-goal orientation on feedback acceptance and employees' creativity. Journal of Leadership & Organizational Studies, 23(1), 82-95.

Spence, J. R., & Keeping, L. M. (2010). The impact of non-performance information on ratings of job performance: A policy-capturing approach. Journal of Organizational Behavior, 31(4), 587-608. doi:10.1002/job.648

Sprinkle, G. B. (2003). Perspectives on experimental research in managerial accounting. Accounting, Organizations and Society, 28(2-3), 287-318.

Sprinkle, G., Williamson, M., & Upton, D. (2008). The effort and risk-taking effects of budget-based contracts. Accounting, Organizations and Society, 33(4), 436-452. doi:10.1016/j.aos.2007.11.001

Stone, D. L., & Stone, E. F. (1985). The effects of feedback consistency and feedback favorability on self-perceived task competence and perceived feedback accuracy. Organizational Behavior and Human Decision Processes, 36(2), 167-185.

Stubbs, K. (2021). Narrative Feedback in Subjective Performance Evaluations: Do Ratings Change the Narrative? Working paper.

Sullivan-Toole, H., Richey, J., & Tricomi, E. (2017). Control and Effort Costs Influence the Motivational Consequences of Choice. Front. Psychol., 8(MAY). doi:10.3389/fpsyg.2017.00675

Tafkov, I. D. (2013). Private and public relative performance information under different compensation contracts. The Accounting Review, 88(1), 327-350.

Tafkov, I. D. (2013). Private and public relative performance information under different compensation contracts.(Report). Accounting Review, 88(1), 327. doi:10.2308/accr-50292

Taggar, S., & Brown, T. C. (2006). Interpersonal Affect and Peer Rating Bias in Teams. Small group research, 37(1), 86-111. doi:10.1177/1046496405284382

Tavoletti, E., Stephens, R. D., & Dong, L. (2019). The impact of peer evaluation on team effort, productivity, motivation and performance in global virtual teams., 25(5/6), 334-347. doi:10.1108/TPM-03-2019-0025

Thaler, R. H. (1999). Mental accounting matters. Journal of Behavioral Decision Making, 12(3), 183-206. doi:10.1002/(SICI)1099-0771(199909)12:3<183::AID-BDM318>3.0.CO;2-F

Thibaut, J. W. and L. Walker (1975). Procedural justice : a psychological analysis, Hillsdale : Erlbaum.

Tice, D. M., Baumeister, R. F., Shmueli, D., & Muraven, M. (2007). Restoring the self: Positive affect helps improve self-regulation following ego depletion. Journal of Experimental Social Psychology, 43(3), 379-384. doi:10.1016/j.jesp.2006.05.007

Tiny Pulse. (2020). These 4 Companies Totally get Employee Recognition. Retrieved from https://www.tinypulse.com/blog/these-4-companies-totally-get-employee-recognition

Tonidandel, S., Quiñones, M. A., & Adams, A. A. (2002). Computer-adaptive testing: The impact of test characteristics on perceived performance and test takers' reactions. Journal of Applied Psychology, 87(2), 320.

Toro-Serey, C., Kane, G. A., & McGuire, J. T. (2021). Choices favoring cognitive effort in a foraging environment decrease when multiple forms of effort and delay are interleaved. Cognitive, Affective, & Behavioral Neuroscience, 1-24.

Toubia, O. (2006). Idea generation, creativity, and incentives. Marketing science, 25(5), 411-425.

Towry, K. L. (2003). Control in a Teamwork Environment: The Impact of Social Ties on the Effectiveness of Mutual Monitoring Contracts. The Accounting Review, 78(4), 1069-1095. doi:10.2308/accr.2003.78.4.1069

Towry, K. L. (2003). Control in a teamwork environment--the impact of social ties on the effectiveness of mutual monitoring contracts. Accounting Review, 78(4), 1069. doi:10.2308/accr.2003.78.4.1069

Trevino, L. K. (1986). Ethical decision making in organizations: A person-situation interactionist model. Academy of management Review, 11(3), 601-617.

Tsai, W. C., Chi, N. W., Grandey, A. A., & Fung, S. C. (2012). Positive group affective tone and team creativity: Negative group affective tone and team trust as boundary conditions. Journal of Organizational Behavior, 33(5), 638-656.

Tzini, K., & Jain, K. (2018). Unethical behavior under relative performance evaluation: Evidence and remedy Human Resource Management, 57(6), 1399-1413. doi:10.1002/hrm.21913

Van den Abbeele, A., Roodhooft, F., & Warlop, L. (2009). The effect of cost information on buyer–supplier negotiations in different power settings. Accounting, Organizations and Society, 34(2), 245-266. doi:10.1016/j.aos.2008.05.005

Van Knippenberg, D. (2000). Work motivation and performance: A social identity perspective. Applied Psychology, 49(3), 357-371.

Venkatesh, R., & Blaskovich, J. (2012). The mediating effect of psychological capital on the budget participation-job performance relationship. Journal of Management Accounting Research, 24(1), 159-175.

Vidal-Salazar, M. D., Cordón-Pozo, E., & José, M. (2016). Flexibility of benefit systems and firms' attraction and retention capacities. Employee Relations.

Vohs, K. D., Baumeister, R. F., Schmeichel, B. J., Twenge, J. M., Nelson, N. M., & Tice, D. M. (2008). Making Choices Impairs Subsequent Self-Control: A Limited-Resource Account of Decision Making, Self-Regulation, and Active Initiative. Journal of Personality and Social Psychology, 94(5), 883-898. doi:10.1037/0022-3514.94.5.883

von Ahn, L., & Dabbish, L. (2004, 2004). Labeling images with a computer game.

Waldman, D. A., Atwater, L. E., & Antonioni, D. (1998). Has 360 degree feedback gone amok? , 12(2), 86-94. doi:10.5465/AME.1998.650519

Wang, L. W. (2017). Recognizing the Best: The Productive and Counterproductive Effects of Relative Performance Recognition. Contemporary Accounting Research, 34(2), 966-990. doi:10.1111/1911-3846.12292

Webb, A. R., Williamson, M. G., & Zhang, Y. (2013). Productivity-Target Difficulty, Target-Based Pay, and Outside-the-Box Thinking. The Accounting Review, 88(4), 1433-1457. doi:10.2308/accr-50436

Weiss, H. M., & Cropanzano, R. (1996). Affective events theory. Research in Organizational Behavior, 18(1), 1-74.

Westbrook, A., & Braver, T. S. (2015). Cognitive effort: A neuroeconomic approach. Cognitive, Affective, & Behavioral Neuroscience, 15(2), 395-415.

Williams, S., & Luthans, F. (1992). The impact of choice of rewards and feedback on task performance. Journal of Organizational Behavior, 13(7), 653-666. doi:10.1002/job.4030130703

Wills, T. A. (1981). Downward comparison principles in social psychology. Psychological Bulletin, 90(2), 245-271. doi:10.1037/0033-2909.90.2.245

Wong-On-Wing, B., Guo, L., & Lui, G. (2010). Intrinsic and extrinsic motivation and participation in budgeting: Antecedents and consequences. Behavioral Research in Accounting, 22(2), 133-153.

Wood, J. V. (1989). What is social comparison and how should we study it? Personality and Social Psychology Bulletin., 22(5), 520-537.

Wood, R. E. (1986). Task complexity: Definition of the construct. Organizational Behavior and Human Decision Processes, 37(1), 60-82. doi:10.1016/0749-5978(86)90044-0

Zappos Insights. Four Peer-to-Peer Ways Zappos Employees Rewarch Each Other. Retrieved from https://www.zapposinsights.com/blog/item/four-peertopeer-ways-zappos-employees-reward-each-other