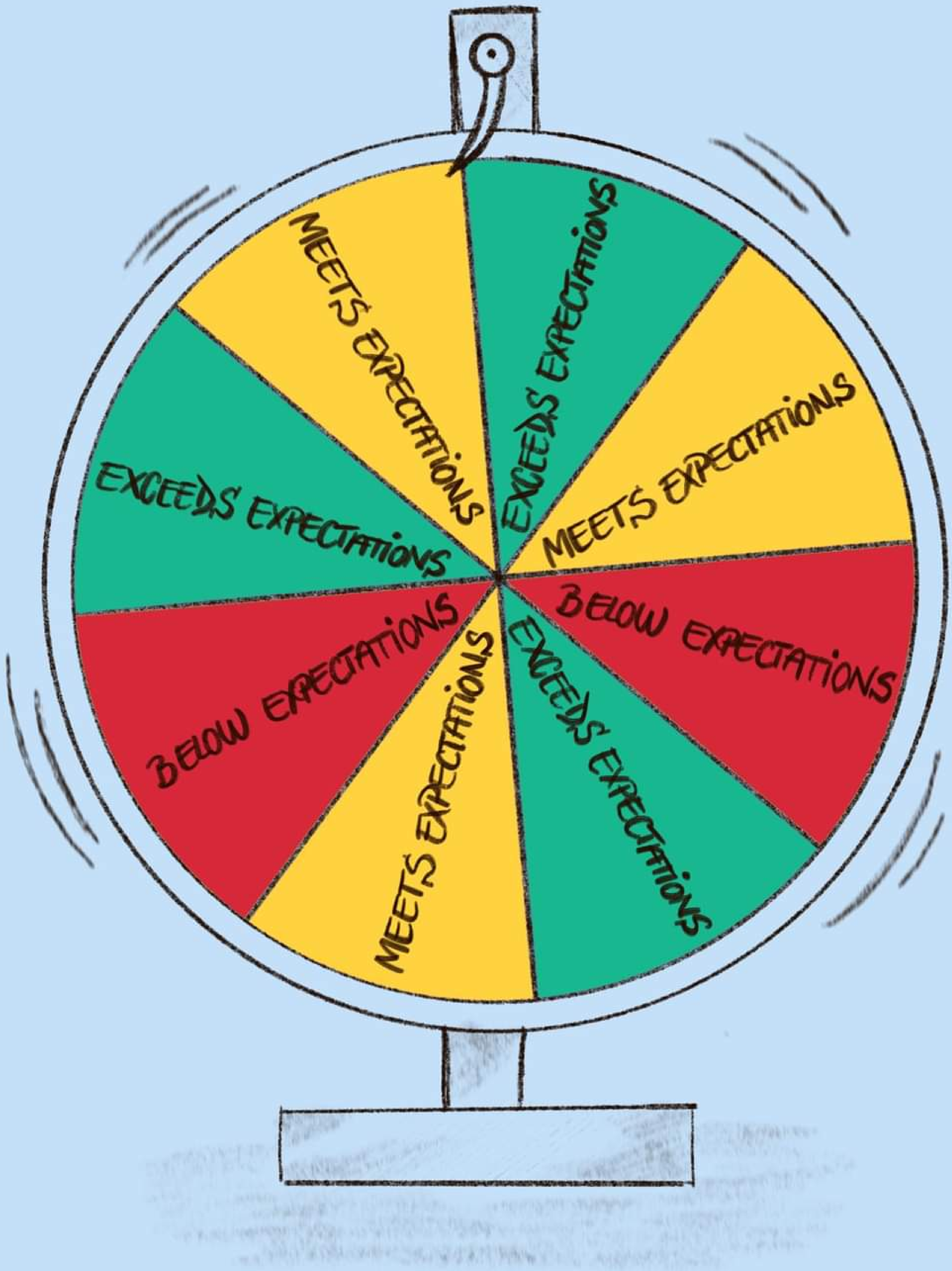
Accounting Information and Discretionary Evaluations



Razvan Ghita

Accounting Information and Discretionary Evaluations

ACADEMISCH PROEFSCHRIFT

ter verkrijging van de graad van doctor
aan de Universiteit van Amsterdam
op gezag van de Rector Magnificus
prof. dr. ir. K.I.J. Maex
ten overstaan van een door het College voor Promoties ingestelde commissie,
in het openbaar te verdedigen in de Agnietenkapel
op woensdag 17 november 2021, te 16.00 uur

door Răzvan-Ştefan Ghiţă
geboren te Caracal

Promotiecommissie

Promotor:	prof. dr. V.S. Maas	Universiteit van Amsterdam
Copromotor:	prof. dr. J.F.M.G. Bouwens	Universiteit van Amsterdam
Overige leden:	dr. J. Choi	University of Wisconsin-Madison
	prof. dr. K.L. Towry	Emory University
	prof. dr. M. van Rinsum	Erasmus Universiteit Rotterdam
	prof. dr. D. Veenman	Universiteit van Amsterdam
	prof. dr. P. Kroos	Universiteit van Amsterdam
	prof. dr. F.H.M. Verbeeten	Universiteit van Amsterdam
	dr. ir. S.P. van Triest	Universiteit van Amsterdam

Faculteit Economie en Bedrijfskunde

Acknowledgements

First and foremost, I am highly grateful to my thesis advisor, Victor Maas. I met Victor at a time in my life when I did not know what to do with my career. Because he was my thesis supervisor, I found out in quick succession what an academic was and that I could become one. During the first years of my PhD, I think I confused Victor many times with my ideas and the way I was communicating them. Through our discussions, I was always amazed by how much autonomy he gave me and how patient he was when I did not use that autonomy well. The first time I spoke during an academic seminar, it was to contradict something Victor said. I knew I was in the right job when all I saw when looking at him was excitement and interest in the particular point I disagreed with. I cannot think of a single instance in which Victor did not support me. Moreover, I can think by many instances in which I was surprised with how much he helped me and how generous he was with his time. Besides being one of the most curious scholars I met, Victor is also an amazing person. Meeting Victor was clearly a case of observable good luck for me and I am grateful for the privilege to have been on this journey with him.

I am also deeply appreciative to the other members of my doctoral committee, Jan Bouwens, Willie Choi, Peter Kroos, Marcel van Rinsum, Kristy Towry, Sander van Triest, David Veenman, Frank Verbeeten. My thesis has greatly benefited from their thoughtful evaluation and feedback at various stages of the manuscript.

I would also love to thank everyone at the Accounting Department at the University of

Amsterdam. UvA was an amazing place to work at. In conversations about work with my friends, they always talked about the organizational impediments (bad bosses, bureaucracy, unhelpful and disinterested colleagues) that prevented them from doing the best work they could do. I never had anything to say during these conversations because I always had all the support I needed at UvA and everyone here wanted me to succeed. I want to thank the people I started the PhD journey with, Jona and Máté. Even though our paths were different, I learned a lot from seeing how different people went through the PhD journey. I would also want to thank the two of the best walk-and-talkers in the world, Pouyan and Ioan. I never truly needed anything from the store, I just wanted to have the walk-and-talk with you.

I am also grateful to the people at Emory. I am grateful to Kristy Towry, for hosting me at Emory University. Kristy taught me much about writing and being a kind and thoughtful person. Going on an academic visit can be a lonely experience but that was not the case for me. Because of the kind effort of everyone, I felt immediately at home at Emory. In this regard, I especially want to thank Jake and Zac for seeing me as a friend. Jake is also a great walk-and-talker.

I appreciate the generous financial support from the University of Amsterdam, the Limperg Institute, and the IMA. This support funded a five-year PhD program, international conference participations, and my academic visit to Emory University. I believe the work that composes this dissertation greatly benefited from comments and suggestions made by workshop participants at Emory University, IE Business School, KU Leuven, Maastricht University, Tilburg University, University of Amsterdam, University of Southern Denmark, the ACMAR 2019 Doctoral Colloquium and 2021 round table discussants, the 2019 ENEAR conference, the 2020 MAS conference, and the 2021 EAA Doctoral Colloquium and conference. I also thank all the kind people in the experimental and management accounting communities. I am always amazed by how much time senior researchers are willing to spend to help junior researchers.

No one has shared more of the hardship of this PhD than my girlfriend (and financial partner!) Bianca. Thank you for all the essential emotional support and thank you for listening to the presentation and reading the papers what must have seemed to you as millions of times. Yet another case of observable good luck to have you in my life. I am always extremely thankful to my friends. I had extraordinary friends before starting the PhD and have only met more amazing people along the way. Thank you all for your support and for helping me take my mind off research. Thank you Victor for the many times you helped with programming.

Finally, I would like to thank my family. Despite never fully understanding what I do, I never doubted that you were as proud of me as you could be. It is the best feeling in the world to know that you are loved unconditionally and I always knew I had that when coming back home. Thank you for giving me independence. I know my moving outside of Romania was difficult and I thank you for supporting this decision. I could not wish for better parents. Thank you to my grandmother who always emphasized the importance of doing well in school. She loves education more than anything despite never having the chance of participating in it herself.

Co-author List

Chapter 3 is joint work with my supervisor Victor Maas. Victor designed the laboratory experiment. He was also involved in designing the case-based experiment, developing the theoretical framework, and positioning the paper. I conducted the empirical analyses and interpreted the results.

External Funding

The research for this doctoral thesis received financial assistance from the Limperg Institute, American Accounting Association Management Accounting Section PhD Student Fund, and Institute of Management Accountants. I am very grateful for their support.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

This dissertation contributes to our understanding of discretionary evaluations, a common and important control mechanism. More precisely, my goal was to better understand what information companies should report to supervisors, and how supervisors should use this information in order to best motivate employees. In chapter 2, I investigate whether increasing reporting frequency affects supervisor evaluation decisions and employee experiential learning in a discretionary evaluation setting. Chapter 3 is co-authored with my supervisor Victor Maas from the University of Amsterdam. We examine whether supervisors' span of control and reporting frequency affect their evaluation decisions. In addition, we examine whether employees anticipate supervisors' reward allocations, and adjust their effort levels based on their supervisor's span of control and the frequency with which their performance is reported. In chapter 4, I examine whether supervisors reward observable good and bad luck in their evaluation decisions, and how this affects employee behavior.

Chapter 2 began with my interest in employee learning. To test my predictions, I developed both an interactive laboratory experiment and an online, case-based experiment. This project highlights my interests in experimental design and in leveraging the competitive advantage of multiple types of experiments. The paper investigates whether increasing reporting frequency affects supervisor evaluation decisions and employee experiential learning in a discretionary evaluation setting. Companies want to motivate employees to explore because the lack of exploration will hamper organizational viability in the long run. Despite this, we know little about how supervisors can motivate exploration in a discretionary evaluation setting. I predict and find that investing effort in unsuccessful exploration results in higher employee bonuses when reporting frequency increases. This is because increasing reporting frequency improves supervisors' ability to distinguish unsuccessful exploration from shirking. Contrary to my prediction, employees do not appear to anticipate this and do not explore more when reporting frequency increases. My results suggest employees can fail to anticipate which actions supervisors will reward, making supervisors less effective at directing employee effort towards desirable actions.

Chapter 3 also reflects my interest in using multiple types of experiments in order to adequately test theoretical predictions. We examine whether supervisors' span of control and reporting frequency affect their evaluation decisions. In addition, we examine whether employees anticipate supervisors' reward allocations, and adjust their effort levels based on their supervisor's span of control and the frequency with which their performance is reported. In an online experiment, we confirm our theory that span of control increases the rewards allocated to top performers and decreases the rewards allocated to the weakest performers. We find no effect of reporting frequency on supervisors' discretionary reward allocations. In a second laboratory experiment, we find no support for our hypotheses that employee effort is affected by span of control and reporting frequency. Our results suggest that widening supervisors' span of control increases evaluation accuracy. However, this increase in accuracy is not sufficient to motivate employees to increase their effort levels.

Chapter 4 reflects my interest in examining the validity of our fundamental assumptions about how supervisors use discretion, and about how they should use discretion in order to best motivate employees. Specifically, the paper examines whether supervisors reward observable good and bad luck in their evaluation decisions, and how this affects employee behavior. Although the controllability principle asserts supervisors should not reward observable luck, I find supervisors reward observable luck because they find it fair to do so. Employees decrease their contribution to company value when supervisors reward observable luck but only after employees learn how supervisors evaluate them through repeated interactions. My results suggest fairness concerns can diminish one of the intended benefits of using subjective evaluations. Specifically, fairness concerns can prevent supervisors from using all available non-contractible information to decrease the weight of luck in employees' compensation.

Chapters 2,3 and 4 tie together in several ways. All chapters examine how providing additional information to supervisors alters their evaluation decisions. All chapters suggest that supervisors use additional information to a lesser extent than would be predicted by normative models (Baiman & Rajan, 1995; Rajan & Reichelstein, 2006). Chapters 2 and 4 suggest that supervisors ignore relevant information about employee behavior because they prefer partially rewarding employees for good and bad luck. Chapter 3 suggest that supervisors ignore relevant information about employee behavior in order to minimize their cognitive load when evaluating employees. All chapters also examine whether employees change their behavior when supervisors evaluate them differently. The results from all chapters suggest that only changing how supervisors evaluate employees is not enough to change employee behavior because employees do not initially anticipate how supervisors will evaluate them. Results from chapters 2 and 4 suggest that employees learn how supervisors will evaluate them through repeated evaluations.

Chapter 2

Reporting Frequency and Learning from Experience

2.1 Introduction

An important dimension of organizational learning and, in turn, organizational performance, is employee on-the-job learning, i.e., experiential learning (Arrow, 1969; March, 1991). To better understand how to motivate employee experiential learning, researchers have investigated what information companies should collect (Dye, 2004) and how this information should be used when designing explicit contracts (Ederer & Manso, 2013; Lee & Meyer-Doyle, 2017; Manso, 2011). However, we know little about what information companies should report to supervisors, how supervisors will use this information in evaluations, and how these evaluations will affect employee experiential learning in a discretionary evaluation setting (Campbell, 2008; Campbell et al., 2011). Since writing comprehensive explicit contracts is sometimes prohibitively costly (Bailey et al., 2011), examining how to motivate experiential learning in a discretionary evaluation setting is important. In this study, I investigate whether increasing the frequency with which employee performance is reported to supervisors affects supervisor evaluation decisions and employee experiential learning in a setting in which supervisors have discretion over employees' rewards.

Examining how reporting frequency affects employee learning can aid organizations in designing their reporting systems. Technological advancements allow companies to relatively cheaply increase reporting frequency to whatever level they consider appropriate (Hecht et al., 2020). Insights into how reporting frequency affects organizational performance are therefore timely because organizations have more flexibility in choosing the reporting frequency that is most likely to maximize profits.

The central tension underlying experiential learning is the choice between the *exploration* of new untested approaches and the *exploitation* of well-known approaches (March, 1991). Many employee decisions can be interpreted as a choice between exploration and exploitation. For example, a salesperson might explore by cold calling potential customers or he might exploit by contacting his existing customer base. Similarly, a manager might explore by contracting with a new supplier or she might exploit by continuing to work with the current suppliers. Because a lack of exploration will hamper organizational viability in the long run (March, 1991), organizations are interested in motivating exploration (Ederer & Manso, 2013).

In a discretionary evaluation setting, employees will only perform an effortful action if they believe that supervisors will infer and reward their effort level (Arnold et al., 2018; Maas et al., 2012; Rajan & Reichelstein, 2006). When choosing between different effortful actions such as exploration and exploitation, employees will consider how likely their supervisor is to infer a high effort level based on the likely outcomes of each action. Hard-working employees will prefer actions that, while possibly suboptimal from a company perspective, are more likely to produce outcomes that inform supervisors about their hard work.

Employees could underinvest in exploration because exploration frequently results in low outcomes and rarely produces a higher outcome than exploitation (Lee & Meyer-Doyle, 2017; March, 1991). Because shirking also produces low outcomes, the supervisor will face an inference problem when observing a low result: did the employee explore or shirk? Consequently, employees who explore have a lower chance that their supervisor will infer their effort level correctly than employees who exploit. As a result, employees could explore less than their company would prefer.

Increasing reporting frequency increases supervisors' ability to distinguish unsuccessful exploration from shirking. Although both unsuccessful exploration and shirking result in low average performance, shirking creates low variability in results while exploration results in high variability (Azoulay et al., 2011; He & Wong, 2004; March, 1991). Increasing reporting frequency allows supervisors to better observe variability in the outcome measure and therefore more confidently interpret low results as either unsuccessful exploration or shirking. Thus, a higher reporting frequency increases the likelihood of supervisors correctly inferring that an employee has invested effort in exploration. Therefore, I expect investing effort in exploration will result in a higher bonus when reporting frequency increases. I further expect employees will anticipate this and explore more when reporting frequency increases.

Notably, prior literature suggests a higher reporting frequency can also decrease employee exploration. First, Campbell et al. (2011) find that a higher reporting frequency increases perceived evaluative pressure which, in turn, reduces employee exploration. However, as the authors note, their data does not allow them to pinpoint whether the decrease in exploration was caused by differences in reporting frequency or by other dimensions of monitoring. Second, Hecht et al. (2020) find that employees are more concerned about producing a low result when reporting frequency is higher. Because exploration frequently results in low results, employees could explore less when reporting frequency is higher. However, employees may be less concerned about presenting low results in a setting where they can learn from experience because the low results could be interpreted as a sign of exploration. Thus, the results of Hecht et al. (2020) do not automatically translate to a setting where employees can learn by exploring. Therefore, the effect of solely manipulating reporting frequency in a setting where employees can learn by exploring is unclear.

I run a interactive lab experiment to answer my research question. Participants were grouped into dyads containing a supervisor and an employee. For ten periods, employees provide costly effort on an abstract task designed to simulate the trade-off between exploration and exploitation. Employees' choices affect company profit which is valuable to the supervisor. The supervisor assigns a bonus to the employee every five periods based on information produced by the reporting system. The supervisor does not pay the bonus out of his or her own pocket. I manipulate reporting frequency at two levels. Supervisors either receive profit reports every period (High Reporting Frequency condition) or aggregated profit reports every five periods (Low Reporting Frequency condition).

Consistent with my theoretical predictions, I find that investing effort in exploration results in a higher employee bonus when reporting frequency increases. However, supervisors do not use all available information to reward in expectation profitable employee actions, preferring instead to partially reward uncontrollable outcomes. When supervisors are better able to observe unsuccessful exploration, supervisors tolerate, instead of reward, unsuccessful exploration. Results from a supplemental case-based experiment provide additional evidence of the effect of reporting frequency on supervisors' evaluations.

Contrary to my theory, I do not find that employees explore more when reporting frequency increases. To investigate why employees did not change their behavior despite supervisors changing their bonus allocations, I examine if employees correctly anticipate how supervisors will evaluate them. I find no evidence that employees anticipate that investing effort in exploration results in a higher bonus when reporting frequency increases. This suggests that employees' uncertainty about how they will be evaluated limits supervisors' ability to direct employee effort towards desirable actions.

Finally, additional analyses indicate that employee risk aversion moderates the effect of reporting frequency on employee exploration. Specifically, reporting frequency does not affect exploration for the most risk-averse employees but increases exploration for the less risk-averse employees. Employees' risk aversion also affects their expectations about how supervisors will reward exploration. Only the less risk-averse employees correctly anticipate that investing effort in exploration results in a higher bonus when reporting frequency increases.

The contribution of my study is threefold. First, I contribute to the management accounting literature by documenting how reporting frequency, a control system choice, affects supervisors' evaluations, and employees' exploratory behavior in a discretionary evaluation setting. Except for a few studies (Campbell, 2008; Campbell et al., 2011), the relationship between management control choices and employee learning has remained unexplored in a discretionary evaluation setting. Second, I contribute to the literature on discretionary evaluations by providing evidence consistent with an additional cost of the uncertainty inherent in discretionary evaluations (Bol, 2008; Luft et al., 2016). My results suggest employees can fail to anticipate which actions supervisors will reward, making supervisors less effective at directing employee effort towards desirable actions. Third, I contribute to the growing behavioral literature that examines the value of providing supervisors with additional information (Hecht et al., 2020; Luft et al., 2016). Agency theory suggests that when companies report additional information, supervisors use all this information to increase evaluation accuracy (Feltham & Xie, 1994; Golman & Bhatia, 2012). I find that when supervisors receive additional information they do not necessarily use all this information to reward effortful employee actions, preferring instead to partially reward uncontrollable outcomes. This suggests that providing additional information to supervisors may be less beneficial than predicted by formal models.

2.2 Theory

2.2.1 Learning from Experience

Organizations learn through a variety of processes, ranging from intentional search to learning from experience (Campbell et al., 2011). Moreover, learning occurs at different organizational levels (e.g. an individual, team, or organizational). I examine employee-level learning in a setting where employees learn from experience. Specifically, I focus on the trade-off between the exploration of new untested approaches and the exploitation of well-known approaches that arises when learning from experience (March, 1991). As March notes, 'the essence of exploitation is the refinement and extension of existing competencies, technologies, and paradigms. The essence of exploration is experimentation with new alternatives' (March, 1991, p. 85).

When learning from experience, an action taken today affects both today's profit (operational outcome) and the amount of knowledge the employee has about the most profitable actions (learning outcome). Exploration and exploitation differ in their expected effect on operational and learning outcomes. In more abstract terms, an employee facing an exploration/exploitation trade-off is operating in a multitasking setting in which exploration and exploitation affect operational outcomes and learning outcomes differently (Hellmann & Thiele, 2011).

When employees engage in exploitation by replicating an action performed in the past, they gain experience and incrementally improve the efficiency of that action (Gupta et al., 2013). Exploitation produces slow, incremental learning concerning the current trajectory. In terms of operational outcomes, because exploitation involves replicating the profit-maximizing action, it usually delivers results that are 'positive, proximate, and predictable' (March, 1991, p. 85).

When employees explore by engaging in a new action for which they do not know the outcome, they learn about the effectiveness of this new action. Exploration produces a more radical type of learning that involves discovering if new trajectories are worth following.

In terms of operational outcomes, exploration produces a distinct pattern of results due to the high ex-ante uncertainty associated with new actions. First, exploration is a variance seeking approach (March, 1991) that produces a higher variability of results (He & Wong, 2004; Taylor & Greve, 2006). By generating a higher variability of results, employees have a higher chance of observing a more successful action (Dye, 2004; March, 1991). Second, because exploration involves not choosing the profit-maximizing strategy based on the current knowledge, it often results in lower outcomes than exploitation.

Previous literature has investigated how to motivate employee exploration using explicit contracts (Ederer & Manso, 2013; Lee & Meyer-Doyle, 2017; Manso, 2011). This literature finds that traditional pay-for-performance contracts are suboptimal for motivating exploration due to the specific return pattern generated by exploration. To motivate exploration, companies need to decrease the risk imposed on employees when exploring (Lee & Meyer-Doyle, 2017) by tolerating early failure and rewarding long term success (Ederer & Manso, 2013; Manso, 2011).

Research examining the exploration/exploitation trade-off spans multiple disciplines (see Mehlhorn et al. (2015)), leading to different perspectives about the exploration/exploitation trade-off. Even within the discipline I draw upon in this study, the organizational learning literature, the assumptions behind exploration and exploitation are heavily debated (Gupta et al., 2013). In this study, I make the following assumptions. First, I follow prior research and examine an environment in which employees take actions and observe outcomes directly attributable to the latest action (Rahmandad, 2008). While environments in which actions produce a delayed response are common and important, learning the causal effects in these environments may be less feasible for individual employees. If actions produce a delayed response, the complexity of the learning problem increases dramatically (Rahmandad, 2008) and humans require many periods to learn the causal relationships in such environments. For example, Gibson (2000) approximates that participants in his study would have required over 1,000 periods to learn the causal model that contained a two-period delay.[1] Second, I follow the recommendations of Gupta et al. (2013) and define both exploration and exploitation as actions that produce knowledge, rather than treating exploration as the only means of producing knowledge.[2]

2.2.2 Reporting Frequency and Exploration

I define reporting frequency as the number and the granularity of reports generated about an employee's performance in a given time interval.[3] By investigating the effects of increasing reporting frequency, I examine the cumulative effect of differences in two reporting dimensions: information aggregation and the time interval between reports. This is an intentional choice because companies likely change both these dimensions simultaneously. More frequent reports naturally involve less information aggregation. Given the recent focus on real-time reporting of performance, less-aggregate reports likely involve the possibility that the supervisor will observe employee performance more frequently. For example, supervisors likely do not need to wait until the end of the month to learn about a salesperson's performance on a given day if the company collects daily information about salespeople's performance. While the effects of solely manipulating information aggregation or the time interval between reports are potentially interesting, they are not the focus of the current study.[4]

Two prior studies suggest increasing reporting frequency decreases exploration. First,

[1]I further discuss this issue when presenting future research opportunities in the 2.6. Discussion and Conclusion section.

[2]Exploitation is a less productive action in settings where this assumption proves untrue. I have no reason to believe that my predictions would change if this is the case.

[3]Reporting frequency is different from feedback frequency. Feedback frequency refers to the frequency with which an employee receives feedback about his or her actions. I do not focus on it in this study and keep it constant across the experimental conditions.

[4]Similar to Hecht et al. (2020), I examine a setting where employees cannot misreport their performance. Instead, the reporting system automatically captures and reports the employees' performance.

Campbell et al. (2011) find that monitoring tightness, a measure that encapsulates reporting frequency, reduces employee exploration. The authors argue that a higher reporting frequency increases perceived evaluative pressure[5] which, in turn, discourages employee exploration. However, Campbell et al. (2011) do not independently examine the effect of reporting frequency on exploratory behavior. Given the nature of field studies, other factors correlated with reporting frequency could be driving the decrease in exploratory behavior. For example, although Campbell et al. (2011) use reporting frequency as one of the criteria to distinguish between tightly monitored and loosely monitored divisions, these divisions may also differ on other dimensions, such as the implicit incentives to refrain from exploration. Given the difficulties in disentangling the effect of reporting frequency with those of other factors using field data, the effect of reporting frequency on exploratory behavior, in isolation, warrants further investigation.

Second, Hecht et al. (2020) find that a higher reporting frequency decreases employee performance by making employees more concerned about presenting low results to supervisors. Thus, increasing reporting frequency should reduce employee exploration because exploration frequently produces low results. However, it is unclear if the conclusions of Hecht et al. (2020) will hold when employees can learn by exploring. Evaluators likely do not learn much about employees' choices and effort levels from the more frequent reports in the study of Hecht et al. (2020) because learning opportunities are limited and performance is largely determined by ability in their design. In contrast, increasing reporting frequency in a setting where employees can learn from experience could allow supervisors to better infer employees' choices and effort levels. Employees could, therefore, be less concerned about presenting low results to supervisors when they can learn by exploring because low results could be interpreted as exploration instead of a lack of ability. Alternatively, the

[5]Evaluative pressure is defined as 'the degree to which salient others are seen as judging rather than enabling one's performance' (Lee et al., 2004, p. 312).

information advantage of increasing reporting frequency could cause employees to explore more despite their increased concern about presenting low results. Therefore, the effect of reporting frequency in a setting where employees can learn from experience is unclear.

2.2.3 Hypothesis Development

The Effects of Higher Reporting Frequency on Supervisors' Evaluations

Supervisors are motivated to use their evaluation decisions to promote employee effort and to direct that effort towards productive actions because higher and better-invested employee effort will likely benefit them (Bol et al., 2016). For example, higher employee effort invested in the correct actions might increase departmental performance, which in turn might increase supervisors' compensation and promotion opportunities. Therefore, supervisors will consider the effects of their evaluation on employee effort and will try to reward effortful actions. Supervisors are more capable of directing employee effort towards specific actions if supervisors can infer the amount of effort invested in that action based on the outcomes they observe (Datar et al., 2001; Feltham & Xie, 1994).

Exploration and exploitation differ in their expected outcomes and in the degree to which these outcomes are informative about employees' effort levels. Due to the focus on experimentation, exploration frequently results in low outcomes, rarely producing a higher outcome than exploitation. Because shirking also produces low outcomes, the supervisor can face an inference problem when observing a low result: did the employee explore or shirk? Thus, exploration can produce outcomes that are uninformative about employees' effort levels because of its high chance of failure and because of the supervisors' inference problem when observing low results.

The bonus allocation patterns of supervisors when observing a low result are likely to vary if supervisors cannot infer how much effort was invested in exploration. Some supervisors may forgo the attempt to base their bonus decision on effort, deciding instead to only consider employee outcomes, the sum of the employees' actions and luck. This strategy would involve the possibility that some employees who exerted high effort are punished for bad luck. Hard-working employees are likely to consider this unfair and possibly retaliate against the supervisor by not exerting effort in future periods (Bol et al., 2016). Wanting to appear fair, other supervisors might still partially reward employees even when they observe low results and are unsure whether the low result was caused by exploration or shirking.

The above logic suggests that the supervisors' inability to distinguish unsuccessful exploration from shirking reduces supervisors' ability to infer the amount of effort invested in exploration and alters supervisors' evaluations. Supervisors will be less likely to reward exploration and punish shirking in their bonus allocation decisions. Therefore, a reporting system that increases the ability of supervisors to distinguish unsuccessful exploration from shirking should translate into higher bonuses for unsuccessful exploration and lower bonuses for shirking.

Increasing reporting frequency should allow supervisors to observe differences in patterns of results produced by exploration and shirking. While both unsuccessful exploration and shirking result in low performance, exploration results in higher variability (He & Wong, 2004; March, 1991). Increasing reporting frequency gives supervisors more information about the variability in performance, increasing their ability to distinguish exploration from shirking. The type of actions that the employees perform is likely to have a strong influence on the variability of performance.[6] Thus, supervisors who observe the variability in performance will use the additional information to infer whether the employees have explored when supervisors observe a low result. As a result, supervisors will be more likely to reward effort invested in unsuccessful exploration and punish shirking in their bonus allocation decisions as reporting

[6]Noise can also affect variability. At some point, if the noise factor has a high enough influence on performance, variability may contain no information about whether an employee explored. However, in such settings, it may not be productive to explore at all because the noise factor makes it difficult to understand the relationships between actions and outcomes (Bohn, 1995).

frequency increases. Formally stated:

H1a: Supervisors award a higher bonus to employees who invest effort in unsuccessful exploration when reporting frequency is higher.

H1b: Supervisors award a lower bonus to employees who shirk when reporting frequency is higher.

The predictions above depend on supervisors using the additional information generated by the increased reporting frequency to reward effort as opposed to outcomes. Some supervisors could exclusively reward outcomes because they find it fair to do so (Ghita, 2021a) or because they cannot anticipate how their evaluations will influence employees' choices (Krishnan et al., 2005). These supervisors will allocate low bonuses to all employees who produced a low result regardless of whether they believe exploration or shirking caused the low results. Nevertheless, I expect that, on average, supervisors will use the additional information generated by the increased reporting frequency to better align the bonuses employees receive with their effort level. This is because, while some supervisors find it fair to reward outcomes, many other supervisors find it fair to reward employee effort (Bol et al., 2015; Chan, 2018; Maas et al., 2012). These supervisors do not need to anticipate how their employees will react to their evaluations (Krishnan et al., 2005) to reward unsuccessful exploration.[7]

The Effects of Higher Reporting Frequency on Employees' Choices

Hard-working employees will prefer to invest their effort in actions that are more likely to produce outcomes that inform supervisors about their hard work. This is because em-

[7]Supervisors may also ignore the additional information about employees' variability in performance generated by increasing reporting frequency because they consider that variability is uninformative about employees' effort levels. If employees anticipate that supervisors will reward variability, employees could generate variability by alternating between shirking and exploiting. Thus, although it is likely true that exploration results in a higher variability than shirking when employees perform the same action consistently (He & Wong, 2004; March, 1991), supervisors may not interpret variability as a signal of exploration because employees can also produce variability in results by strategically choosing when to shirk. I present arguments and evidence against this possibility in the 2.5. Supplemental Experiment section.

ployees' bonuses are less influenced by factors outside their control if supervisors can infer employees' effort levels. Therefore, employees prefer performing actions that are likely to produce informative outcomes since such actions require them to bear less risk (Cadsby et al., 2019).[8]

When reporting frequency is low, employees may anticipate supervisors' difficulty in inferring whether unsuccessful exploration or shirking caused a low result. Given that reaction to low results is likely to vary across supervisors, employees will be more uncertain about how supervisors will evaluate low results when reporting frequency is lower. Because unsuccessful exploration could be less likely to be rewarded, employees will need to bear more risk when exploring. Employees who are not protected against the risk of exploration through tolerance for failure explore less (Ederer & Manso, 2013; Lee & Meyer-Doyle, 2017; Manso, 2011). Thus, employees will be less likely to explore when reporting frequency is lower. Moreover, because low results are less likely to be punished, shirking will become a more attractive action. Thus, employees will be more likely to shirk when reporting frequency is lower.

Employees' exploitation efforts will likely also be affected by supervisors' difficulty in distinguishing unsuccessful exploration from shirking. Even if employees understand that exploration is the optimal action from the perspective of company and supervisor, the high chance that supervisors will observe effort invested in exploitation will make exploitation a safer and possibly more attractive action for employees. Thus, employees will be more likely to exploit when reporting frequency is lower. Formally stated, the hypothesis is:

H2: Employees explore more, exploit less, and shirk less when reporting frequency is higher.

[8]This argumentation assumes that most employees are risk-averse. This assumption is likely true given that most people are risk-averse (Holt & Laury, 2002) and companies are unlikely to be able to select only risk-seeking employees through their selection processes because these selection processes are designed to achieve multiple objectives.

This hypothesis is not without tension. If supervisors base their evaluations on output and not on employee effort, employees will bear the entire risk of exploration regardless of reporting frequency. Employees understand supervisors are influenced by luck in their evaluations (Brazel et al., 2016; Brownback & Kuhn, 2019). Therefore, if employees correctly anticipate that supervisors base their evaluations on output, employee choices will not be influenced by reporting frequency.

Even if supervisors exclusively reward employee effort, employee might nevertheless not change their behavior. This is because employees might not anticipate how supervisors will evaluate them (e.g. employees might overestimate how much supervisors are influenced by output). This lack of mutual understanding between supervisors and employees about what supervisors value in their evaluation is a documented cost of discretionary evaluations (Bol, 2008; Gibbons & Henderson, 2012; Luft et al., 2016), and may mute the effect of increasing the reporting frequency.

2.3 Research Method

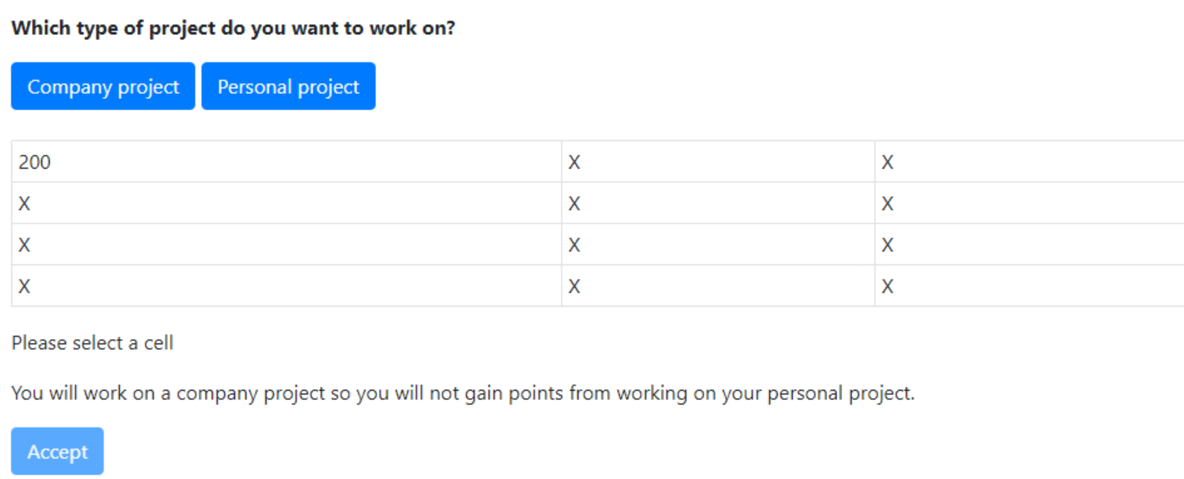
2.3.1 Task Design

I randomly assigned participants to assume the role of either a supervisor or an employee. One supervisor and one employee formed a company. In each of the ten periods, employees provided costly effort on an abstract task designed to simulate the trade-off between exploration and exploitation. Employees' effort choices and a noise factor determined company profit in each period. Company profit is valuable to supervisors. Supervisors decided how much of a fixed bonus pool to allocate to the employee in their company every five periods. Thus, supervisors evaluated employees twice during the task. Supervisors received reports containing their company's profit information. In the High Reporting Frequency condition, supervisors received profit reports every period, while in the Low Reporting Frequency condition, supervisors received aggregated profit reports every five periods.

Employee Task

Every period, employees chose between working on their own personal project, a familiar company project, and an unfamiliar company project. Working on their *own personal project* generated a private benefit of 50 points for employees and did not contribute to company profit. Working on a company project (either familiar or unfamiliar) did not generate private benefits for the employee, but contributed to company profit. Employees could choose what kind of company project to work on from the cells of a table like the one displayed in Figure 2.1. Cells marked with a number represent working on a familiar company project and cells marked with X represent working on an unfamiliar company project.

Figure 2.1: Example of Employees' Decision Screen



This figure shows the screen employees saw when making their decision in period one.

In the first period, the table contained one familiar project that contributed 200 points to company profit and eleven unfamiliar company projects that contributed an unknown number of points to company profit. When the employee chose to work on a *familiar company project* (a cell marked with a number), the employee contributed to company profit with the number marked on that cell. To capture the learning effects of exploitation, the points contributed to company profit from working on a familiar project increased in the remaining periods. Specifically:

$\text{Profit}_{\text{new}} = \text{Profit}_{\text{old}} * 110\%$, for the first three periods in which the project is chosen.

$\text{Profit}_{\text{new}}$ was rounded to the nearest integer. After the third period in which an employee chose the same project, the profit associated with that project remained constant. Working on the same project only produced improvements for a limited number of periods because repeating the same activity is likely to produce only a limited amount of learning (Gupta et al., 2013).

When employees chose to work on an *unfamiliar company project* (a cell marked with an X), they discovered that project's contribution to company profit (the X is replaced by a number). The newly discovered project contribution became the employee's contribution to company profit in that period. The project became a familiar company project in all remaining periods. All employees began the task with the company projects displayed in 2.1.

Table 2.1: Company Projects in Period One

200 (familiar)	*400 (unfamiliar)*	0 (unfamiliar)	0 (unfamiliar)
60 (unfamiliar)	60 (unfamiliar)	80 (unfamiliar)	80 (unfamiliar)
100 (unfamiliar)	100 (unfamiliar)	120 (unfamiliar)	120 (unfamiliar)

The parameters in the table captured the theoretical outcomes of exploration and exploitation presented in the 2.2. Theory section. Specifically, working on a familiar company project (the operationalization of exploitation) resulted in a medium performance. Working on an unfamiliar company project (the operationalization of exploration) was rarely successful and produced variable results when exploration was unsuccessful.

To preclude participants from forming different expectations about the distribution of the table based on their own experience, I informed both employees and supervisors about the underlying distribution of the grid. In the real world, employees and supervisors likely also hold beliefs about the outcomes of exploration, exploitation, and shirking, and those beliefs are likely in line with the theoretical predictions on which I base the parameters.[9]

Supervisor Task

Supervisors received reports containing profit information and decided how much bonus their employees would receive (Figure 2.2). Supervisors awarded the bonus twice, once after the fifth period and again after the tenth period. Supervisors had full discretion in determining the bonuses, which ranged between zero and 500 points.

The bonuses were paid from a fixed bonus pool. Although some prior studies use a variable bonus pool that increases as a function of an objective performance measure such as company profit (Bailey et al., 2011; Fisher et al., 2005; Maas et al., 2012), I used a fixed bonus pool because a variable bonus pool would have implicitly transferred some of the risk of failure onto the employees, thus adding a confounding factor.

Payoffs

I used the following payoff functions (in points, with each point worth 0.01 euro) to determine participants' compensation:

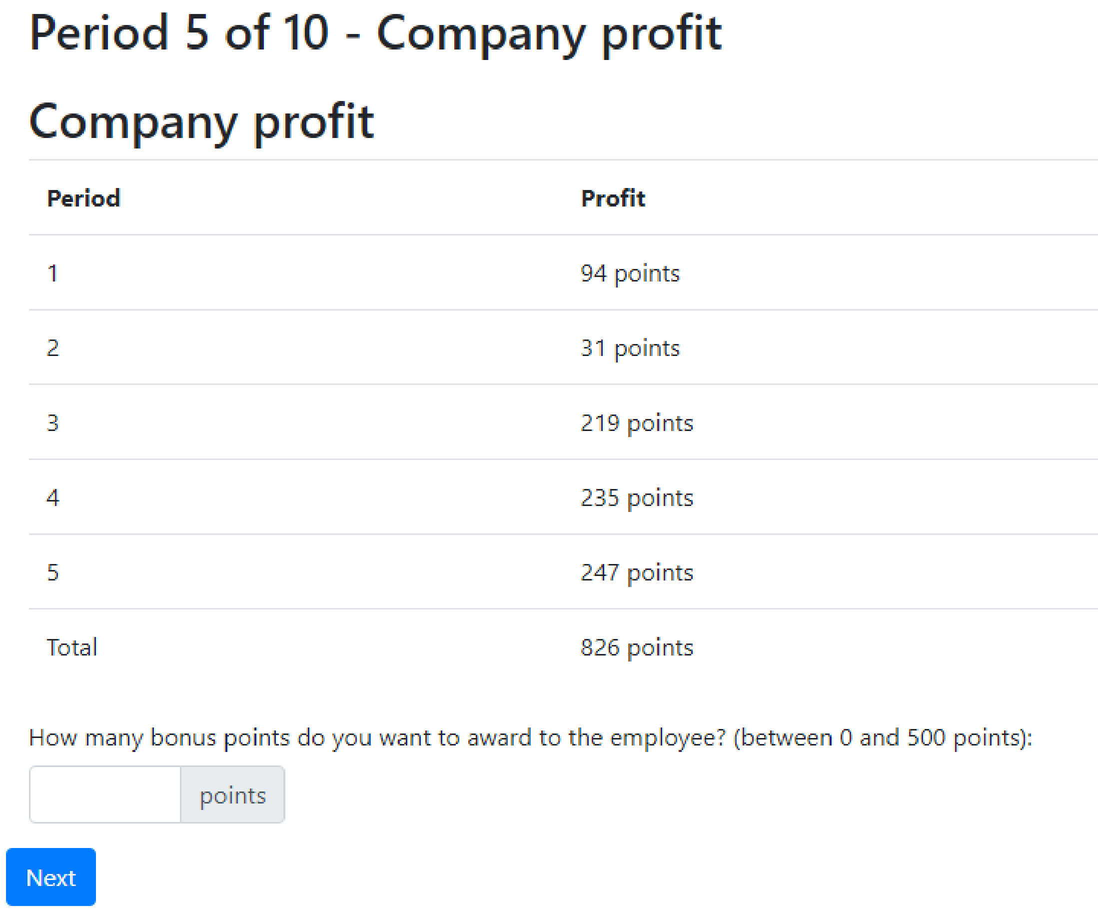
Employee: First Bonus + Second Bonus + 50 * (Number of periods of working on their own personal project)

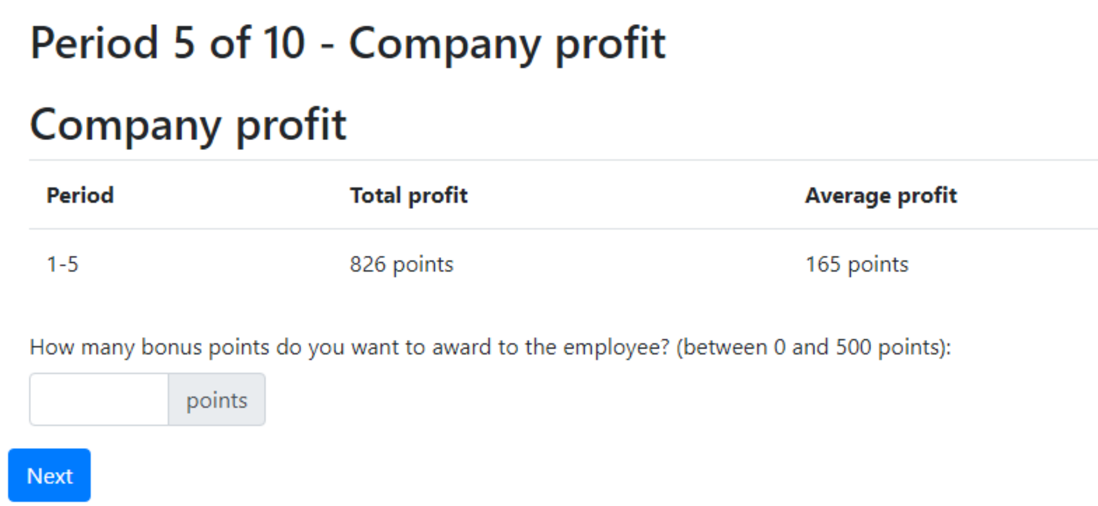
Supervisor: 50% * (aggregate company profit during the last five periods)

Testing my hypotheses requires a setting where supervisors wanted to promote exploration (i.e., for employees to work on unfamiliar company projects). Otherwise, no employees would have chosen to work on unfamiliar company projects, transforming the setting into a single-action agency problem. Therefore, to make exploration a desirable action, I based su-

[9] In addition to employees' project choices, a noise factor ranging from zero to twenty points also affected company profit. Both employees and supervisors knew about the existence and distribution of this noise factor.

Figure 2.2: Examples of Supervisors' Evaluation Screen





This figure shows the screen supervisors saw when making their evaluation decision after period five for both Reporting Frequency conditions.

pervisors' compensation on the company's profit in the last five periods. This compensation scheme did not punish supervisors for the low results of unsuccessful exploration but allowed them to benefit in case exploration was successful (Ederer & Manso, 2013; Manso, 2011). Supervisors are more likely to want to promote employee exploration when working under this compensation scheme as compared to a traditional pay-for-performance compensation scheme in which their payoff would depend on employee performance in all ten periods.

2.3.2 Employees' Optimal Choices from the Company's Perspective

In this section, I discuss the sequence of project choices that maximizes expected company profit. This sequence entails that employees always work on company projects, and never work on their personal projects. In a *single-period setting*, the optimization strategy entails choosing the type of project that maximizes expected company profit. Since the expected value of the familiar company project is higher than the expected value of an unfamiliar company project (200 points versus 102 points[10]), a profit-maximizing employee will work on the familiar company project.

When the employees make choices for more than one period, they have a chance to learn: the employees can work on unfamiliar projects in the hope of discovering the project that produces 400 points and choosing it in all remaining periods. Therefore, calculating the expected outcome of each type of company project no longer leads employees to the optimal strategy when the employees make choices for multiple periods because it underestimates the value of working on unfamiliar company projects.

A profit-maximizing employee would always work on unfamiliar projects in the first periods (they would postpone working on familiar company projects until the later periods of the task) because discovering the 400-point project in the early periods is more profitable than discovering it in the later periods. Therefore, a possible strategy is to commit to a

[10]The expected value of working on a unfamiliar company project is calculated by dividing the sum of profits for all unfamiliar projects by the total number of available unfamiliar company projects.

maximum number of periods in which to work on unfamiliar company projects (hereafter, a switching point). If the employees would discover the 400-point project before the switching point, then the employees would work on that project for all remaining periods. If the employees would not discover the 400-point project before the switching point, then the employees would stop working on unfamiliar company projects and choose the best familiar project available (the 200-point project) for the remaining periods. I present expected company profit for each possible switching point in Table 2.2 and provide the Python program used to calculate this in Appendix 1.[11] Thus, as shown in Table 2.2, a risk-neutral company prefers that employees choose to work on unfamiliar company projects for a maximum of three periods, then select the most profitable familiar company project.

Table 2.2: Expected Outcome for Each Switching Point

Switching point	0	1	2	3	4	5	6	7	8	9	10
Expected outcome	2,524	2,565	2,600	2,628	2,623	2,616	2,607	2,596	2,593	2,596	2,603

This table presents the expected outcome (rounded to the nearest integer) for each switching point. A switching point is the maximum number of periods that an employee decides to work on unfamiliar company projects. If the employee discovers the 400-point project before the switching point, then the employee works on that project in all remaining periods. If the employee does not discover the 400-point project before the switching point, then the employee stops working on unfamiliar company projects and chooses the best familiar company project available (the 200-point project) in the remaining periods.

Given the difficulties of calculating the optimal switching point, participants likely would

[11] This strategy is naïve because the employees partially commit to a switching point and only depart from the switching point if they find the 400-point project. However, a more effective strategy entails determining the optimal switching point in a manner that goes beyond whether the 400-point project is discovered. For example, it is more profitable to keep working on unfamiliar company projects after the zero-point project is discovered than when a 120-point project is discovered because the expected profit of the remaining unfamiliar company projects is higher when the zero-point project is discovered. Calculating the optimal switching point using this strategy is more difficult and computationally expensive given how large the decision tree becomes. However, because the naïve strategy underestimates how profitable it is to work on unfamiliar company projects, the expected outcomes of the naïve strategy serve as a minimum switching point.

not have been able to calculate the optimal switching point for themselves. Thus, participants would have likely reached different conclusions about which employee choices result in the highest payoff for supervisors and the company. Therefore, to avoid such heterogeneous conclusions and increase the power of my tests, I presented several hypothetical choices in the instructions explaining the optimal order of choices for an employee who is interested in maximizing company profit. These examples also included hypothetical payoffs for the supervisors and the company given different employee choices.

2.3.3 Procedures

Participants received an initial set of instructions (e.g. no communication, do not use phones) before moving to the computer lab. I provided participants with a hardcopy of the instructions and gave them fifteen minutes to read the instructions before starting the next phase of the experiment.

The computerized portion of the experiment was programmed in OTree (Chen et al., 2016). The computer experiment started with a quiz. Participants needed to correctly answer every quiz question before proceeding with the next phase of the experiment. Then, all participants completed five practice periods in which they assumed the role of an employee. Their performance during the practice periods did not affect their final payoff. Then, I randomly assigned participants to either the employee or supervisor role. Participants maintained their role throughout the session.

Participants performed the main task of the experiment four times during a session. I randomly selected one of the four tasks as the payoff task, and participants received their payoffs from this task. After each task, I randomly re-matched employees and supervisors to form new companies. Therefore, an employee was unlikely to interact with the same supervisor more than once during a session, and vice versa.

Each of the four tasks consisted of ten periods. In each period, employees chose between working on a familiar company project, working on an unfamiliar company project, and working on their personal project. After every choice, employees received summary information that included their contribution to company profit, the amount of noise affecting company profit, the change in available company projects for the remaining periods, and the amount of private benefits they gained. The summary information also included the employee's complete history of project choices and outcomes.

After the employee's project choices in periods five and ten, supervisors assigned a bonus to the employees in their companies. Employees indicated how much bonus they expected to receive before being informed about their supervisors' decision. At the end of every task, supervisors indicated their expectations about the frequency with which employees in their companies chose each type of project in the ten periods of that task.

Finally, participants completed a post-experimental questionnaire. This questionnaire included items for measuring participants' perspective-taking abilities and risk attitudes. The questionnaire also contained items intended to provide insights into participants' decision-making process during the experiment.

2.3.4 Participants

I recruited participants from the participant pool of CREED, an experimental economics laboratory at the University of Amsterdam. In total, 124 individuals participated in six sessions. To ameliorate potential session effects, each session contained both experimental conditions. Participants' age in years ranged from 18 to 58, with a mean of 21.51 and a median of twenty years. In total, 51 participants (41.1%) were male and 70 participants (56.45%) were female. Three participants did not disclose their gender. Participants reported an average work experience of 1.18 years and 84 participants (67%) indicated economics or business as their main area of study. Participants received a €5 participation fee in addition to the payoff they earned during the experiment. On average, participants earned a total of €14.76 for about one hour of their time. The average payoff of participants in the role of supervisor (€14.51) is similar to that of participants in the role of employee (€15.01).

2.3.5 Dependent Variables

I examine supervisors' bonus allocation decisions as the main dependent variable related to supervisors' behavior. I examined three dependent variables related to employee behavior. First, I measured exploration as the decision to work on an unfamiliar company project in a period. This operationalization satisfies the definition of exploration because choosing to work on an unfamiliar company project involves experimenting with a new alternative for which the outcome is unknown. Second, I measured exploitation as the decision to work on a familiar company project in a period. The operationalization satisfies the definition of exploitation because choosing to work on a familiar company project involves using existing knowledge (how much profit does a project produce) to extract value and refining it in order to generate greater marginal returns in future periods. Third, I measured shirking as the decision to work on the personal project in a period. Baiman (1982) defines effort as a construct that is controllable by the employee, creates disutility for the employee, and increases expected output for the company. The operationalization of shirking (lack of effort) is appropriate because working on the personal project is a choice the employees make, creates personal benefits for the employees, and does not contribute to company profit.

For the main tests of the hypotheses, I examine supervisor bonus decisions and employee choices in the first five periods of each task. I do so for two reasons. First, my theory assumes that exploration has a high chance of failure. This assumption is more likely to be true in the first five periods.[12] Second, I assumed supervisors want to award bonuses that motivate employees to exert effort and that employees anticipate this. This is more likely to be the case before the first evaluation period because both supervisors and employees knew they would continue to interact with each other after the evaluation. Therefore, supervisors knew that they could motivate better choices from employees in the last five periods if they awarded an

[12] Table 2.3 presents the individual and cumulative probability of discovering the 400-point project depending on the number of periods in which an employee chose to work on unfamiliar company projects. I inform participants about the cumulative probability of discovering the 400-point project.

appropriate bonus in the first evaluation period. In contrast, both supervisors and employees knew that they would be unlikely to interact again after the second evaluation period. Thus, supervisors were less likely to use the bonus in the second evaluation to motivate employees.

I measured participants' risk attitudes in the post-experimental questionnaire using the instrument developed by Holt & Laury (2002) (oTree implementation by Holzmeister (2017)). My measure, *RiskAversion*, is the total number of safe choices from the instrument of Holt & Laury (2002).[13]

Table 2.3: Probabilities of Discovering the 400-Point Project

	Probability per period		
Period	Formula	Result	Cumulative probability
1	1/11	9.09%	9.09%
2	1/10	10.00%	18.18%
3	1/9	11.11%	27.27%
4	1/8	12.50%	36.36%
5	1/7	14.29%	45.45%
6	1/6	16.67%	54.55%
7	1/5	20.00%	63.64%
8	1/4	25.00%	72.73%
9	1/3	33.33%	81.82%
10	1/2	50.00%	90.91%

This table presents the probability of discovering the 400-point project for each period of working on unfamiliar company projects and the associated cumulative probability. The cumulative probability is calculated using the formula $1 - \prod_{1}^{k=i}(1 - Prob(k))$.

[13] I also measured perspective-taking ability using the Perspective-Taking Scale developed by Davis (1980) because the effect of reporting frequency on employee/supervisor behavior may be stronger for participants who are better at perspective taking. Additionally, understanding how differences in perspective-taking ability affect employee behavior could add to our understanding of agency theory (Foss & Stea, 2014). I find no meaningful interactions between perspective taking and reporting frequency when analyzing employee/supervisor behavior so I do not further discuss perspective taking.

2.4 Results

2.4.1 Randomization Check and Descriptive Statistics

Participants in the two treatments do not differ with respect to age, gender, work experience, number of experimental sessions attended in the past month, perspective-taking ability, and risk preferences. Results of a multiple linear regression do not reveal an association between the measured characteristics and assignment to one of the conditions ($F(6, 117) = 0.84$, $p > 0.10$, $R^2 = 0.04$). In addition, participants are also similar in terms of the same measured characteristics between the two roles ($F(6, 117) = 0.78$, $p > 0.10$, $R^2 = 0.04$). These analyses suggest that random assignment was successful.

In total, I collected 2,480 employee-period observations ([124 participants/2 roles] x 10 periods x 4 tasks) and 496 supervisor-evaluation observations ([124 participants/2 roles] x 2 evaluation periods x 4 tasks).[14] Table 2.4 presents the bonuses allocated by supervisors in the first evaluation period conditional on the number of periods of exploration and shirking across the two conditions. Table 2.5 and Figure 2.3 present the proportion of employees who chose to explore, exploit, and shirk, as well as the proportion of employees who successfully explored across the two conditions and across the five periods.

Before testing the hypotheses I check whether the manipulation was effective. Results from manipulation checks included in the post-experimental questionnaire suggest that most employees were attentive to what information their supervisors received. I asked participants what information did the reports presented to supervisors contain: 'company profit for each period separately', or 'only the aggregate and average profit for five periods'. 22 of the 31 employee-participants in the Low Reporting Frequency condition (70.97%) and 24 of the 31

[14] In five tasks, a technical error caused the computers to not update the table from the previous task. Thus, five participants started a new task with the action table from the previous task. I eliminate the data generated by these tasks from my analyses (50 employee-period observations and ten supervisor-evaluation observations). It is unlikely that this technical issue affects the conclusions of the study because I am still able to analyze approximately 98% of the data (243 out of the total of 248 tasks).

Table 2.4: Supervisors' Bonus Allocations Depending on Employees' Choices

Panel A - The Effect of Unsuccessful Exploration on Bonus

		Period of Exploration					
Condition		0	1	2	3	4	5
Low Frequency	mean	397	333	386	386	325	254
	sd	130	199	135	137	143	180
	n	15	6	9	10	9	40
High Frequency	mean	280	450	250	264	400	290
	sd	222	71	132	170	104	158
	n	14	2	3	7	7	43

Panel B - The Effect of Shirking on Bonus

		Periods of Shirking					
Condition		0	1	2	3	4	5
Low Frequency	mean	315	448	431	338	-	458
	sd	163	65	94	189	-	102
	n	94	11	7	4	0	6
High Frequency	mean	370	372	288	200	275	44
	sd	151	139	199	0	106	79
	n	98	9	5	1	2	6

This table presents the bonus awarded by supervisors in the first evaluation period depending on how many periods they choose to explore (Panel A) and shirk (Panel B) across the two conditions.

Table 2.5: Employee Choices by Period

Panel A - First Five Periods

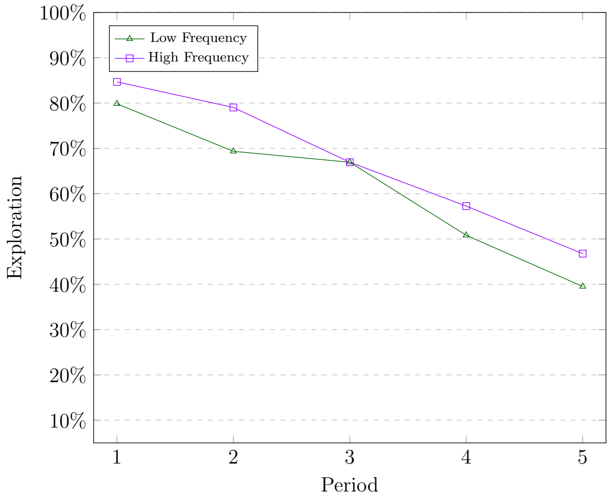
	Low Frequency			High Frequency		
Period	Explore	Exploit	Shirk	Explore	Exploit	Shirk
1	79.84%	12.10%	8.06%	84.68%	9.68%	5.65%
2	69.35%	18.55%	12.10%	79.03%	12.90%	8.06%
3	66.94%	21.77%	11.29%	66.94%	21.77%	11.29%
4	50.81%	38.71%	10.48%	57.26%	31.45%	11.29%
5	39.52%	48.39%	12.10%	46.77%	40.32%	12.90%
T	61.29%	27.90%	10.81%	66.94%	23.22%	9.84%

Panel B - Last Five Periods

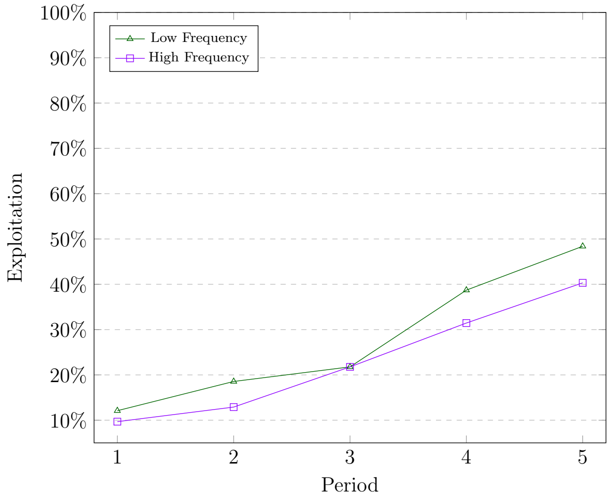
	Low Frequency			High Frequency		
Period	Explore	Exploit	Shirk	Explore	Exploit	Shirk
6	33.06%	48.39%	18.55%	24.19%	58.06%	17.74%
7	25.81%	59.68%	14.52%	14.52%	66.94%	18.55%
8	12.90%	69.35%	17.74%	14.52%	66.94%	18.55%
9	16.13%	66.94%	16.94%	12.10%	70.16%	17.74%
10	8.87%	67.74%	23.39%	8.06%	71.77%	20.16%
T	19.35%	62.42%	18.23%	14.68%	66.77%	18.55%

This table presents the proportion of employees who chose to explore (chose to work on a company project with an unknown outcome), exploit (chose to work on a company project with a known outcome), and shirk (chose to work on a personal project) across the two conditions and across the ten periods. To test the hypothesis related to employee choices, I analyze the first five periods. The employee choices for the first five periods are presented in Panel A. The employee choices for the last five periods are presented in Panel B for descriptive purposes.

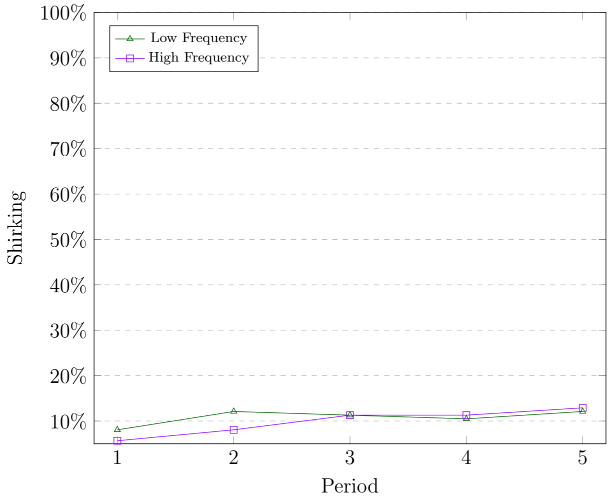
Figure 2.3: Employee Behavior per Period



(a) Exploration per Period



(b) Exploitation per Period



(c) Shirking per Period

employee-participants in the High Reporting Frequency condition (77.42%) answered this question correctly.[15] Since all participants also passed the understanding test in advance of the experiment, I test the hypotheses with the full sample.[16]

2.4.2 Hypotheses Tests

The Effects of Higher Reporting Frequency on Supervisors' Evaluations

H1a predicts that supervisors award a higher bonus to employees who invest effort in unsuccessful exploration when reporting frequency is higher. To test H1a, I examine the interaction between *HighReportingFrequency* (equal to 1 for the High Reporting Frequency condition and 0 for the Low Reporting Frequency condition) and *NumPeriodsExploration* (the number of periods in which the employees worked on an unfamiliar company project in the first five periods) on *FirstBonus* (the bonus decision of the supervisor after period five). A positive interaction coefficient would support H1a and suggest that an additional period of exploration (as compared to an additional period of exploitation or shirking) increases the bonus more (or decreases the bonus less) in the High Reporting Frequency condition as compared to the Low Reporting Frequency condition.

Specifically, I estimate the following linear regression:[17]

FirstBonus = *NumPeriodsExploration* + *HighReportingFrequency* + *NumPeriodsExploration* x *HighReportingFrequency*

To focus on the probability of rewarding unsuccessful exploration, I examine the tasks

[15] The likelihood of employee-participants failing the manipulation check is not significantly different between conditions. The coefficient of *HighReportingFrequency* in a logit regression that examines the chance of failing the manipulation check is $\beta = 0.34$, $z = 0.58$, $p > 0.10$, two-tailed.

[16] It is also possible that the wording or the positioning of the manipulation check question confused participants. 26 of the 31 supervisor-participants in the low frequency condition (83.87%) and 28 of the 31 supervisor-participants in the high frequency condition (90.32%) answered the manipulation check correctly. It is unlikely that supervisor-participants in the High Reporting Frequency condition did not notice or remember the reporting frequency manipulation given that they needed to click a button in every period after they received a report in order to advance the experiment.

[17] I use period-by-period data in all regressions reported in the paper (rather than one observation per participant). Therefore, standard errors are clustered on the supervisor or employee level to control for multiple observations within a participant.

in which employees did not discover the 400-point project. Table 2.6, Column 1 reports the results of this regression. Consistent with H1a, the interaction coefficient is marginally significant and positive (β = 29.50, t = 1.87, p < 0.10, two-tailed), indicating an additional period of unsuccessful exploration increases the bonus more in the High Reporting Frequency condition compared to the Low Reporting Frequency condition. This result supports H1a.

Through their bonus allocations, supervisors make exploration a more advantageous action for employees when reporting frequency is higher.[18] This, however, does not mean that supervisors choose to exclusively reward effort (and ignore outcomes) when they become better at inferring employees' effort choices.[19] Instead, exploration becomes more profitable for employees because supervisors tolerate, rather than reward, unsuccessful exploration in the High Reporting Frequency condition and punish unsuccessful exploration in the Low Reporting Frequency. The coefficient of *NumPeriodsExploration* in Table 2.6, Column 1 is significant and negative (β = -28.48, t = - 2.89, p < 0.01, two-tailed) indicating that supervisors punish unsuccessful exploration in the Low Reporting Frequency and the interaction term (β = 29.50, t = 1.87, p < 0.10, two-tailed) is high enough to only offset this negative effect. Indeed, when I only analyze data from tasks in which employees unsuccessfully explore in all five periods, the coefficient of *HighReportingFrequency* is not statistically significant (β = 23.18, t = 0.71, p > 0.10, two-tailed). These results suggest that supervisors do not use all the information available to them to reward profitable (in expectations) and effortful employee actions, preferring instead to partially reward uncontrollable outcomes.

H1b predicts that supervisors award a lower bonus to employees who shirk when report-

[18]Supervisors in the Low Reporting Frequency condition do not offset their punishment of unsuccessful exploration by rewarding successful exploration more. Successful exploration increases employee bonus to a similar extent between the two reporting frequency conditions. In an untabulated analysis, I find that the interaction between *Found400*, a dummy variable that takes the value 1 if the employee has discovered the 400-point project and 0 otherwise, and *HighReportingFrequency* is not statistically significant (β = 49.08, t = 0.97, p > 0.10, two-tailed).

[19]I verify that supervisors are better at inferring employees' effort choices when reporting frequency is higher in the 2.4.3. Additional Analyses section.

Table 2.6: Effect of Employee Choices on Supervisors' Bonus Decisions and Expected Bonus

VARIABLES	(1) FirstBonus	(2) FirstBonus	(3) FBE	(4) FBE
HighReportingFrequency	-116.68*	51.12	-60.80	34.16
	(66.81)	(33.11)	(82.42)	(28.65)
NumPeriodsExploration	-28.48***		-18.47	
	(9.86)		(11.93)	
HRF*NumPeriodsExploration	29.50*		15.34	
	(15.74)		(19.59)	
NumPeriodsShirking		31.42***		6.12
		(10.53)		(13.81)
HRF*NumPeriodsShirking		-86.60***		-36.32
		(13.25)		(24.39)
Constant	411.37***	323.88***	368.30***	332.08***
	(35.28)	(25.24)	(49.51)	(21.30)
Observations	165	243	165	243

This table indicates the results of regressions analyzing supervisors' bonus decisions and employees' expectations about the bonus decisions. The dependent variables are: FirstBonus in Columns 1 and 2 (the bonus decision of the supervisors after period five) and ExpectedFirstBonus (EFB) in Columns 3 and 4 (employees' expectation about the bonus decision of the supervisors after period five). The independent variables are: HighReportingFrequency (equal to 1 for the High Reporting Frequency condition and 0 for the Low Reporting Frequency condition); NumPeriodsExploration (the number of periods in which the employees worked on an unfamiliar company project in the first five periods); and NumPeriodsShirking (the number of periods in which the employees worked on a personal project in the first five periods). To examine how unsuccessful exploration affects bonuses, I analyze the tasks in which employees did not discover the 400-point project. Therefore, the number of observations is lower in the regressions of Columns 1 and 3.
***, **, * denote statistical significance at the 1%, 5%, 10% level. Robust standard errors clustered at the participant level (supervisors for Columns 1 and 2 and employees for Columns 3 and 4) are presented in parenthesis. The hypotheses related to supervisor evaluation decisions predict that investing effort in unsuccessful exploration results in a higher employee bonus when reporting frequency is higher, and that shirking results in a lower employee bonus when reporting frequency is higher. The interaction effect of HighReportingFrequency and NumPeriodsExploration from Column 1 and of HighReportingFrequency and NumPeriodsShirking from Column 2 are consistent with the hypotheses. However, the results of Columns 3 and 4 are not consistent with the assumption that employees anticipate supervisors' bonus allocation decisions.

ing frequency is higher. To test H1b, I examine the interaction between *HighReportingFrequency* and *NumPeriodsShirking* (the number of periods in which the employees worked on the personal project in the first five periods) on *FirstBonus*. A negative interaction coefficient would support H1b and suggest that an additional period of shirking (as compared to an additional period of exploitation or exploration) decreases the bonus more (or increases the bonus less) in the High Reporting Frequency condition as compared to the Low Reporting Frequency condition. Specifically, I estimate the following linear regression:

$$FirstBonus = NumPeriodsShirking + HighReportingFrequency + NumPeriodsShirking \times HighReportingFrequency$$

Differently from the test for H1a, I did not drop tasks in which the employee has successfully explored when testing the effect on shirking on bonuses because supervisors in the Low Reporting Frequency condition are likely unaware that the employee is no longer facing an exploration-exploitation trade-off. Therefore, the employees could use this ambiguity to shirk after they have successfully explored.

Table 2.6, Column 2 reports the results of this regression. Consistent with H1b, the interaction coefficient is statistically significant and negative (β = -86.60, t = -6.54, p < 0.01, two-tailed), indicating an additional period of shirking decreases the bonus more in the High Reporting Frequency condition compared to the Low Reporting Frequency condition. This result supports H1b.

The Effects of Higher Reporting Frequency on Employees' Choices

H2 predicts employees explore more, exploit less, and shirk less when reporting frequency is higher. To test H2, I examine the effect of *HighReportingFrequency* on the following three variables:

- *Exploration*, equal to 1 if the employee works on an unfamiliar company project in a period, and 0 otherwise

- *Exploitation*, equal to 1 if the employee works on a familiar company project in a period, and 0 otherwise

- *Shirking*, equal to 1 if the employee works on a personal project in a period, and 0 otherwise

Specifically, I estimate the following logit model:

$Y = \mathit{HighReportingFrequency} + \mathit{Period}$

Where Y is either *Exploration*, *Exploitation* or *Shirking*

Since the predictions are based on the assumption that employees are facing an exploration-exploitation trade-off, I drop all employee-period observations that occur after the employee had discovered the 400-point project in a given task, as the employee no longer faces an exploration-exploitation trade-off in these periods.[20] Similar to the test of H1b, I do not drop these employee-period observations when examining the effect on shirking. As explained in the 2.3.5. Dependent Variables section, I drop all observations after the fifth period.

Table 2.7 reports the results of these analyses. *HighReportingFrequency* does not have a statistically significant effect on *Exploration* (β = 0.41, z = 0.89, p > 0.10, two tailed), *Exploitation* (β = -0.40, z = -0.78, p > 0.10, two tailed) or *Shirking* (β = -0.11, z = -0.2, p > 0.10, two tailed). Thus, I do not find support for H2.

[20]Because the dependent measures are censored after an employee discovers the 400-point project, I test whether censoring affects the results. Discarding the censored observations introduces a downward bias in the coefficient of *HighReportingFrequency*. To control for censoring, I can use Inverse Probability Weighting (IPW) (Wooldridge, 2007) which gives a higher weight to observations that had a higher likelihood of being censored. The likelihood of being censored is calculated based on the chance of discovering the 400-point project given how many periods the employee has explored in a given task. I can use IPW because my design allows me to precisely calculate the change of being censored in each period and employees are similar to each other because of random assignment (Wooldridge, 2007). I find qualitatively similar results when I correct for censoring using IPW (results untabulated). Therefore, I conclude censoring does not significantly affect my results.

Table 2.7: Effect of Reporting Frequency on Exploration, Exploitation and Shirking

VARIABLES	(1) Exploration	(2) Exploitation	(3) Shirking
HighReportingFrequency	0.41	-0.40	-0.11
	(0.46)	(0.51)	(0.58)
Period	-0.30***	0.36***	0.13**
	(0.05)	(0.06)	(0.06)
Constant	1.75***	-2.57***	-2.50***
	(0.36)	(0.42)	(0.51)
Observations	1,046	1,046	1,215

This table indicates the results of regressions analyzing employee choices. The dependent variables are: Exploration in Column 1 (equal to 1 if the employee works on an unfamiliar company project in a period, and 0 otherwise), Exploitation in Column 2 (equal to 1 if the employee works on a familiar company project in a period, and 0 otherwise), and Shirking in Column 3 (equal to 1 if the employee works on a personal project in a period, and 0 otherwise). The independent variables for all three regressions are: HighReportingFrequency (equal to 1 for the High Reporting Frequency condition and 0 for the Low Reporting Frequency condition); Period represents the period of the observation (ranging from 1 to 5). The number of observations differs in Column 3 because the subsample that is most relevant for testing the effect of HighReportingFrequency on Shirking includes periods in which the employees had discovered the 400-point project.
***, **, * denote statistical significance at the 1%, 5%, 10% level. Robust standard errors clustered at each employee level are presented in parenthesis. H2 predicts HighReportingFrequency will affect all three dependent variables. HighReportingFrequency does not significantly affect any of the dependent variables, providing no support for H2.

2.4.3 Additional Analyses

The previous tests suggest employees did not change their behavior when their supervisors received more frequent reports. In this section, I first test the assumption that led me to predict that employees will change their behavior when their supervisors received more frequent reports. I examine (1) if supervisors are more likely to correctly identify employee actions when reporting frequency increases, (2) if employees correctly anticipate that their supervisors will be more likely to reward unsuccessful exploration and punish shirking when reporting frequency increases, and (3) if employees behave differently in the last task after they had a chance to learn how supervisors evaluate them. Afterward, I investigate if employees' risk-aversion moderates the relationship between reporting frequency and employee exploration. Finally, I examine the supervisors' bonus allocations in the second evaluation period.

Supervisors' Beliefs about Employee Choices

In my theory development, I assumed supervisors are more capable of observing the effort invested in exploration as reporting frequency increases. The experimental design allows me to test this assumption. In the post-experiment questionnaire, supervisors indicated their agreement with the statement 'I could determine how many periods the employees worked on unfamiliar company projects based on the profit figures reported to me' on a Likert Scale from 1 (strongly disagree) to 5 (strongly agree). Supervisors in the High Reporting Frequency condition indicated stronger agreement (mean = 3.80) than supervisors in the Low Reporting Frequency condition (mean = 3.10) ($t = 3.07$, $p < 0.05$, two-tailed). This suggests supervisors believe they are more capable of observing effort invested in exploration when reporting frequency is higher.

At the end of every task, supervisors answer the question: 'how many periods do you think the employee chose to work on an unfamiliar company project?' For each of the four tasks, I calculated the absolute difference between (1) supervisors' expectations about the number of exploration periods (the response to the previous question) and (2) the actual number of exploration periods chosen by employees. Then, I regress this difference on *HighReportingFrequency*. Results (untabulated) reveal this difference is smaller when reporting frequency increases ($\beta = -0.83$, $t = -2.05$, $p < 0.05$, two-tailed) suggesting that supervisors are more capable of observing effort invested in exploration as reporting frequency increases.

Employees' Beliefs about Bonus Allocations

In my theory development, I assumed employees are more likely to believe that investing effort in unsuccessful exploration results in a higher bonus when reporting frequency is higher. I also assumed that employees are more likely to believe that shirking results in a lower bonus when reporting frequency is higher. To test these assumptions, I analyze employees' expectations about how much bonus they will receive in the first evaluation period. The regression specifications and sample are similar to those reported in Table 2.6, Columns 1 and 2, except that the dependent measure is *ExpectedFirstBonus* (*EFB*), which captures employees' expectations about how much bonus they will receive in the first evaluation period. Table 2.6, Column 3 reports the results of the regression that analyzes the expected effect of unsuccessful exploration on bonuses. The interaction coefficient is not statistically significant (β = 15.34, t = 0.78, p > 0.10, two-tailed). Table 2.6, Column 4 reports the results of the regression that analyzes the expected effect of shirking on bonuses. The interaction coefficient is not statistically significant (β = -36.32, t = -1.49, p > 0.10, two-tailed). These results do not support the assumption that employees are more likely to expect effort invested in unsuccessful exploration to result in a higher bonus and shirking to result in a lower bonus when reporting frequency is higher.

Employee Choices in the Last Task

Employees explore more and exploit less in the last task when reporting frequency is higher. This is consistent with the idea that employees learn that exploration is more advantageous when reporting frequency is higher (H1a). When I run the regressions used to test H2 (the models presented in Table 2.7) with a subsample of observation from the last task, I find that *HighReportingFrequency* has a marginally significant and positive effect on *Exploration* (β = 0.85, z = 1.52, p < 0.10, one-tailed), a significant and negative effect on *Exploitation* (β = -1.12, z = -1.67, p < 0.05, one-tailed) and no significant effect on *Shirking* (β = -0.10, z = -0.15, p > 0.10, one-tailed).[21]

[21] One-tailed tests are appropriate because the tension underlying these tests is whether or not employees learn that effort invested in unsuccessful exploration in the first five periods results in a higher bonus when reporting frequency is higher. Given that supervisors reward unsuccessful exploration more in the first five periods when reporting frequency is higher, I cannot find a reason why employees would conclude that unsuccessful exploration results in a lower bonus when reporting frequency is higher.

The Moderating Effect of Risk Aversion on Employee Choices

Employees' risk aversion might influence the effect reporting frequency has on exploration. My theory suggests that increasing reporting frequency increases the probability that exploration will be rewarded regardless of whether luck determines that exploration is successful. Risk-averse employees do not want their bonus to be influenced by luck so they will explore more when reporting frequency increases. This line of argument suggests that my prediction will only apply to risk-averse employees. However, it is also possible that extremely risk-averse employees believe their supervisors do not want them to explore. People's own risk-aversion strongly influences their beliefs about other's risk aversion (Chakravarty et al., 2011) so extremely risk-averse employees may believe their supervisors will not reward exploration (a relatively risky action) because they do not want them to explore. Moreover, because exerting effort involves accepting uncertainty when supervisors have discretion over employees' rewards (Bol, 2008), extremely risk-averse employees may not exert effort regardless of the information reported to supervisors. Given these possible interactions with employees' risk aversion, investigating if risk aversion moderates the effect of reporting frequency on exploration is informative.

I investigate if there is a significant interaction between *RiskAversion* and *HighReportingFrequency* on *Exploration*. Recall that the *RiskAversion* measure was obtained by adding the total number of safe choices from the instrument developed by Holt & Laury (2002). The interaction between *RiskAversion* and *HighReportingFrequency* is marginally significant (β = -0.46, t = -1.71, p < 0.10, two-tailed) indicating that changing reporting frequency increases exploration less when employees are more risk-averse. When I run the regressions used to test H2 (the models presented in Table 2.7) with a subsample of employees who report levels of risk-aversion that are lower or equal to the sample median (39 employee-participants, 62.90% of the total sample of 62 employee-participants), I find that *HighReportingFrequency* has a significant and positive effect on *Exploration* (β = 1.28, z = 2.08, p < 0.05, two-tailed), a marginally significant and negative effect on *Exploitation* (β = -1.21, t = -1.79, p < 0.10, two-tailed) and no significant effect on *Shirking* (β = -0.13, t = -0.16, p > 0.10, two-tailed). Consistent with the idea that employees' own risk-aversion affects their beliefs about whether their supervisor will reward risky actions, I find that less risk-averse employees believe that their supervisors will reward unsuccessful exploration. When I run the regressions used to test whether employees believe unsuccessful exploration will be rewarded (the models presented in Table 2.6, Column 3) with a subsample of employees who report levels of risk-aversion that are lower or equal to the sample median, I find that the interaction between *NumPeriodsExploration* and *HighReportingFrequency* is marginally significant (β = 40.97, t = 1.78, p < 0.10, two-tailed) indicating that these employees believe that effort invested in unsuccessful exploration results in a higher bonus when reporting frequency is higher.

Bonus Allocations when Exploration Decreases Supervisors' Payoffs

I analyze supervisors' bonus allocation in the second evaluation period of a task. Supervisors' incentives are different in the second evaluation period as compared to the first evaluation period. First, unsuccessful exploration decreases supervisors' payoffs more in the periods before the second evaluation period as compared to the periods before the first evaluation period. This is because the supervisors' payoffs only depended on company profit in the last five periods of each task. Second, because of the anonymous rematching after each task, supervisors are unlikely to interact with the same employee after the second evaluation period. Thus, supervisors are less motivated to allocate bonuses that they expect will increase employee effort in future interactions. Instead, prior literature suggests that supervisors are guided by their fairness concerns in such situations (Maas et al., 2012).

I find that, in the second evaluation period, supervisors are more likely to punish unsuccessful exploration when reporting frequency is higher. I analyze the effect of employees' choices in the last five periods of the task on supervisors' bonus allocations in the second evaluation period using the same models as the ones used to test H1a and H1b (the models presented in Table 2.6, Columns 1 and 2). I find that the interaction between *HighReportingFrequency* and *NumPeriodsExploration* is significant and negative (β = -53.89, t = -2.45, $p < 0.05$, two-tailed) indicating an additional period of exploration decreases the bonus more in the High Reporting Frequency condition compared to the Low Reporting Frequency condition.[22]

Supervisors punish an effortful employee action, unsuccessful exploration, more when reporting frequency increases and they become better at inferring employees' effort choices. Previous literature suggests that providing additional information to supervisors decreases the risk imposed on employees when they perform risky and effortful actions partially because fairness concerns drive supervisors to reward effortful actions (Arnold et al., 2018; Chan, 2018; Maas et al., 2012). This result suggests that providing additional information to supervisors can also increase the risk imposed on the employees and that supervisors' incentives are likely an important determinant of how supervisors use this additional information.

2.5 Supplemental Experiment

An important assumption in my theory development is that supervisors interpret variability in employee results as a signal of employee exploration. This assumption may not necessarily be true if supervisors believe employees act strategically. If employees anticipate that supervisors will reward variability, employees could generate variability by alternating between shirking and exploiting. Thus, although it is likely true that exploration results in a higher variability than shirking when employees perform the same action consistently (He & Wong, 2004; March, 1991), supervisors may not interpret variability as a signal of exploration because employees can also produce variability in results by strategically choosing when to

[22]The interaction between *HighReportingFrequency* and *NumPeriodsShirking* remains significant and negative (β = -34.28, t = -2.28, $p < 0.05$, two-tailed) indicating an additional period of shirking decreases the bonus more in the High Reporting Frequency condition compared to the Low Reporting Frequency condition.

shirk.

Despite this, I expect most supervisors will interpret the variability in results as employee exploration. First, the logic outlined above assumes individuals have full strategic thinking capabilities, that is, they fully incorporate others' strategy into their decisions. Research finds that most people do not have full strategic thinking capability (Camerer et al., 2004; Cardinaels et al., 2018). Therefore, supervisors may not realize that employees would intentionally induce variability in their results to appear as if they have unsuccessfully explored. Second, because supervisors have private incentives to give high bonuses to employees (Ahn et al., 2010; Bol et al., 2016), supervisors may be willing to put more weight on any piece of information that justifies giving the employee a high bonus (Du et al., 2018; Moers, 2005). Thus, supervisors could increase the bonus of employees when they observe variability in performance even though they understand that variability is not an unequivocal signal of employee exploration. Third, employees may not have a strong incentive to fake exploration because they are uncertain whether supervisors will (1) interpret variability as exploration and (2) reward exploration (Bol, 2008; Gibbons & Henderson, 2012; Luft et al., 2016).

The lab experiment allowed employees to generate variability by strategically shirking. However, in my lab design, employees could perform a single action in a given period. This partially limited employees' ability to induce variability in results when not exploring. Therefore, to provide additional evidence that supervisors interpret variability in results as a signal of employee exploration, I performed an additional case-based experiment. This experiment was also intended to provide additional insights into how variability in performance influences supervisors' bonus allocations.

2.5.1 Design

Participants assumed the role of a regional manager and received performance information about two salespeople. Participants' main task was to indicate how likely they thought each salesperson was to explore and to allocate a bonus to each salesperson. I manipulated reporting frequency between subjects by providing participants with monthly sales reports in the High Reporting Frequency condition and aggregated sales reports for six months in the Low Reporting Frequency condition. Within subjects and nested within the High Reporting Frequency condition, I manipulated the variability in the results of each salesperson.

Participants read a hypothetical case scenario about a company (CoffeeAndGo) that sells beverages from semi-mobile stands. Participants needed to evaluate two of the salespeople working for CoffeeAndGo. Participants learned that because the sales network of CoffeeAndGo was dispersed, salespeople were not directly monitored and could freely decide how many hours to work and wherein their sales area to place their semi-mobile stand. Salespeople could decide if they would explore by moving their stand to a new location to possibly discover a more profitable selling location, or exploit by keeping the stand in their usual selling location. Employees could also shirk by working fewer hours.

Participants were informed about how CoffeeAndGo expected exploration to affect performance. Participants learned that salespeople could maximize their long-term profit by discovering the most profitable location in which to place the stand. The case informed participants that salespeople could only discover the profit potential of a new location by selling in that location. Participants also learned that it usually took a salesperson between two and three months to discover the profit potential of a new location and that most new locations had lower profit potential than the usual selling location of a salesperson.

In their role as regional managers, participants needed to allocate bonuses to two salespeople under their supervision, Bob Stevens and Mark Jonson. The case informed participants that they were expected to maintain a long-term supervisory relationship with Bob and Mark and that regional managers had full discretion in allocating salespeople a mid-year bonus between $0 and $1,000. To ensure that participants were not motivated to underestimate performance in order to use the unallocated bonus for their own interest, they were informed that money that was not rewarded to the salespeople was redistributed to the company bonus pool. Before making the bonus decision, participants indicated how likely they thought each salesperson was to have explored.

The case indicated that CoffeeAndGo provides regional managers with performance reports containing profit information for the current year and, for comparison, profit information for the same period in the previous year. Participants learned that both salespeople performed worse than they had in the previous year. Bob generated $22,000 this year as compared to $24,000 in the previous year and Mark generated $22,500 this year as compared to $24,500 in the previous year. In the High Reporting Frequency condition, participants could observe that the variability in Bob's results was higher than the variability in Mark's results. Specifically, Bob performed relatively well for four out of the six months (his average performance in these four months was similar to Bob's average performance in the previous year, $4,000) and poorly for two consecutive months (his average performance in these two months was $3,000). Mark had low variability in his results and constantly performed worse than in the previous year (his average performance per month was $3,750 as compared to $4,083 in the previous year).

The case indicated that there was no evidence of changes in selling conditions (e.g. demand, competitors) between this year and the last and that profit largely depended on how many hours salespeople worked and on the location of the semi-mobile stand. Participants learned that the two salespeople they needed to evaluate worked in different sales areas of the same city and that they faced similar selling conditions.

The primary purpose of this supplemental experiment was to investigate if supervisors interpret variability in employee performance as a signal of exploration. Therefore, the main dependent variable relates to participants' assessment of how likely each employee was to have explored. To capture this, participants indicated on a seven-point Likert scale how much they agreed with the following sentences: 'I believe that Bob [Mark] changed his location during the first six months of the current year'. To capture how observing variability influenced supervisors' assessments of the likelihood of employee exploration, I examine whether participants were more likely to believe that Bob explored than they were to believe that Mark explored. Given that in the High Reporting Frequency condition, participants could observe that Bob had higher variability in performance than Mark, a higher difference between Bob and Mark in the High Reporting Frequency condition than in the Low Reporting Frequency condition would provide support for the idea that supervisors interpret variability in employee performance as a signal of exploration. To provide further insight into how reporting frequency affects supervisors' bonus allocations, I also analyzed supervisors' bonus decisions.

I used Prolific to recruit 76 participants who reported having supervisory duties at work. Prolific is a crowdsourcing platform that allows researchers to collect data from specific target populations. When participants sign up to Prolific, they fill out a survey about themselves in which they report, among other things, their work responsibilities (Palan & Schitter, 2018). Researchers are then able to prescreen participants based on these reported characteristics. Using this feature, I recruited participants who reported having supervisory duties at work and authority to give instructions to at least one employee. To avoid misunderstanding of the case due to language issues, I required participants to reside in the United States or the United Kingdom and to speak English as a first language. Participants who fulfilled these criteria could follow a link to a Qualtrics instrument. All the data were collected within one session. Participants were randomly assigned to one of the two conditions by Qualtrics.[23]

Participants read the case description and needed to correctly answer eight understanding questions to proceed to the main decision of the study and to receive the participation

[23] Participants' characteristics do not differ between the two treatments with respect to age, gender, work experience, employment status (full-time worker or part-time worker), supervisory experience, number of subordinates and self-reported optimism, risk preferences, and propensity to trust others. Results of a multiple linear regression do not reveal an association between the measured characteristics and assignment to one of the conditions ($F(9, 66) = 0.98$, $p > 0.10$, $R^2 = 0.11$).

fee. Participants who failed to correctly answer all the understanding questions reread the instructions and could attempt to answer the understanding questions again. Participants who failed to correctly answer all understanding questions during their second attempt were not allowed to participate in the study and were not compensated. I informed the participants about this procedure at the beginning of the instructions.

In total, 125 participants attempted to complete the study. Out of these, I eliminated 28 participants because they attempted to complete the study on a mobile device although I requested through the Prolific platform that participants needed to use a desktop device. Three participants did not attempt to answer the understanding questions. During the first attempt, 31 participants failed to correctly answer all the understanding questions. Out of these 31 participants, six did not attempt to answer the attention questions again and twelve failed to correctly answer the attention questions during their second attempt. Therefore, the main sample contains 76 participants.

Out of the 76 participants, 33 (43.42%) were male and 43 (56.58%) were female. On average, participants were 35.72 years old. Participants reported an average of 16.29 years of work experience and 5.57 years of supervisory experience. Participants reported having an average of 4.14 (direct and indirect) subordinates at work.

After the main decision, participants completed a post-experimental questionnaire. This questionnaire contained items that collected information about participants' work experience and provided insights into participants' decision-making process during the experiment. Participants who successfully completed the study received $1.70 for an average of approximately nine minutes of their time (resulting in an average hourly rate of $11.45).

I included one item in the post-experimental questionnaire to check whether participants remembered the frequency with which they received reports about employee performance. I asked participants to indicate what the profit reports presented to them contained: 'the profit generated for each month separately', or 'only the aggregated profit generated for all the six months'. In both conditions, all participants correctly indicated the frequency of the performance reports.

2.5.2 Results

Table 2.8 presents descriptive statistics for each of the two experimental conditions. Panel A presents participants' assessment of how likely each employee was to explore and the differences in assessment between the two employees, i.e., the assessment of whether Bob (participants could observe that this employee had high variability in results in the High reporting frequency condition) explored minus the assessment of whether Mark (participants could observe that this employee had low variability in results in the High reporting frequency condition) explored, indicated by Difference. This difference in assessment between Bob and Mark is higher in the High Reporting Frequency condition (mean = 2.00, sd = 2.20) than in the Low reporting frequency condition (mean = 0.11, sd = 0.39). The difference is statistically significant according to Welch's t-test (unequal variances t-test)[24], $t(39.27) = 5.19$, $p < 0.01$ (one-tailed).[25] Participants are more likely to believe Bob explored in the High Reporting Frequency condition (mean = 5.47, sd = 1.17) than in the Low Reporting Frequency condition (mean = 4.92, sd = 1.36) according to Welch's t-test, $t(72.50) = 1.89$, $p < 0.05$ (one-tailed) and are less likely to believe that Mark explored in the High Reporting Frequency (mean = 3.47, sd = 1.67) condition than in the Low Reporting Frequency condition (mean = 4.82, sd = 1.45) according to Welch's t-test, $t(72.53) = 3.74$, $p < 0.01$ (one-tailed). These results support the idea that supervisors interpret variability in employee performance as a signal of exploration.

[24] I use Welch's t-test instead of the more widely used Student's t-test because Welch's t-test produces more robust results when the underlying population variances are unequal and does not reduce power when the underlying population variances are equal (Ruxton, 2006). Results remain qualitatively similar if I use Student's t-test.

[25] To examine whether this result was driven by inexperienced supervisors, I perform the t-test on a subsample of participants that report more than two years of supervisory experience (n = 43). The results are consistent with the full-sample analysis ($t(19.28) = 4.33$, $p < 0.01$, one-tailed) indicating that the result is not driven by inexperienced participants.

These differences in assessments influence participants' bonus decisions. Table 2.8, Panel B presents the bonuses awarded across the two conditions. The difference in bonus between the two employees is higher in the High Reporting Frequency condition (mean = 92, sd = 218) as compared to the Low Reporting Frequency condition (mean = -33, sd = 57). Welch's t-test, $t(42.06) = 3.39$, $p < 0.01$ (two-tailed) confirms that the difference is statistically significant. Participants allocate a higher bonus to Bob in the High Reporting Frequency condition (mean = 569, sd = 270) than in the Low Reporting Frequency condition (mean = 454, sd = 267) according to Welch's t-test, $t(73.99) = 1.86$, $p < 0.10$ (two-tailed). This indicates that, in a setting where employees can learn by exploring, supervisors who observe variability in employees' results increase employees' bonuses. Participants allocate similar bonuses to Mark in the High Reporting Frequency condition (mean = 478, sd = 277) as compared to the Low Reporting Frequency condition (mean = 487, sd = 276) according to Welch's t-test, $t(74.00) = 0.15$, $p > 0.10$ (two-tailed). This indicates that supervisors who observe low variability in employees' results do not decrease employees' bonuses.

Finally, I find that some supervisors do not change their evaluations when reporting frequency increases. These supervisors base evaluations on employees' outcomes instead of employees' effort choices. In the post-experimental questionnaire, participants indicated how much their bonus decision was influenced by rewarding salespeople for their effort and how much their bonus decision was influenced by rewarding salespeople for the profit they generated. Similar to an analysis performed by Maas et al. (2012), I classify participants as prioritizing profit over effort in their evaluations if they indicated that they were more influenced by profit than by effort. Based on this criterion, I classify 30 out of the 76 participants (39.47%) as supervisors who prioritize rewarding profit over effort. Table 2.8, Panel C presents how this classification influences bonuses across the two conditions. In the High Reporting Frequency condition, the difference in bonuses between Bob and Mark is not significantly higher than 0 when analyzing participants who prioritize profit over effort,

Table 2.8: Results - Supplemental Experiment

Panel A – Beliefs about Employee Exploration				
Condition		Salesperson		
		Bob	Mark	Difference
Low Frequency	mean	4.92	4.81	0.11
	sd	1.36	1.45	0.39
	n	38	38	38
High Frequency	mean	5.47	3.47	2
	sd	1.18	1.67	2.22
	n	38	38	38
Panel B – Bonus				
Condition		Salesperson		
		Bob	Mark	Difference
Low Frequency	mean	454	487	-33
	sd	267	276	57
	n	38	38	38
High Frequency	mean	569	478	92
	sd	270	277	218
	n	38	38	38
Panel C – Bonus Conditional on Supervisors' Priority				
Condition		Profit over effort		
		No	Yes	Total
Low Frequency	mean	-28	-40	-33
	sd	59	55	57
	n	25	13	38
High Frequency	mean	150	20	92
	sd	269	100	218
	n	21	17	38

This table presents the descriptive statistics regarding three variables for each of the two experimental conditions. The variable of interest in Panel A is participants' beliefs about whether each employee explored. Specifically, participants indicated their agreement on a seven-point Likert scale to the following sentence "I believe that [Bob or Mark] changed his location during the first six months of the current year" (strongly disagree - strongly agree). The variable of interest in Panel B is the allocated bonus. The participants allocated a bonus between $0 and $1,000 to each of the two salespeople. The variable of interest in Panel C is the difference between the bonus of Bob and Mark. Panel C splits the date depending on participants' self-reported considerations when making the bonus decisions. Specifically, in the post-experimental questionnaire, participants indicated how much their bonus decision was influenced by employee effort and by profit (not at all - a great deal). If participants indicated that they were more influenced by profit than by effort, I classify them as prioritizing profit over effort. The experiment was designed such that Bob and Mark generated similar aggregated profit. Participants in the High Reporting Frequency could observe that the variability in Bob's results is higher than the variability in Mark's results.

(mean = 20, sd = 24, $t(16) = 0.80$, $p > 0.10$, two-tailed) and is significantly higher than 0 when analyzing participants who do not prioritize profit over effort (mean = 150, sd = 59, $t(20) = 2.55$, $p < 0.05$, two-tailed). Supervisors who prioritize profit over effort in their evaluations do not incorporate their inferences about employee exploration in the bonus decisions. When analyzing participants who prioritize profit over effort, the bonus differences between Bob and Mark are statistically similar when participants considered Bob was more likely to explore than Mark (mean = 8, sd = 29) and when they do not consider this (mean = -16, sd = 19) according to Welch's t-test, $t(19.55) = 0.66$, $p > 0.10$ (two-tailed). When analyzing participants who do not prioritize profit over effort, the bonus differences between Bob and Mark are higher when participants considered Bob was more likely to explore than Mark (mean = 158, sd = 65) than when they do not consider this (mean = -21, sd = 13) according to Welch's t-test, $t(19.42) = 2.70$, $p < 0.05$ (two-tailed).

2.6 Discussion and Conclusion

I investigate whether reporting frequency affects supervisor evaluation decisions and employee experiential learning in a discretionary evaluation setting. I find that investing effort in exploration results in a higher employee bonus when reporting frequency increases. However, when supervisors are better able to observe unsuccessful exploration, they tolerate, instead of reward, unsuccessful exploration. Results from a supplemental case-based experiment provide additional evidence for the effect of reporting frequency on supervisors' evaluations. I find no evidence that employees explore more when reporting frequency increases. Employees likely do not change their exploration behavior because they do not anticipate that investing effort in exploration results in a higher bonus when reporting frequency increases.

My contribution to research is threefold. First, I contribute to the management accounting literature by examining if reporting frequency, a control system choice, influences the exploratory behavior of employees in a discretionary evaluation setting. Except for a few studies (Campbell, 2008; Campbell et al., 2011), the relationship between management control choices and employee experiential learning has remained unexplored in a discretionary evaluation setting. I find a higher reporting frequency allows supervisors to better observe exploration and increases the bonus employees obtain by exploring. However, employees do not explore more when reporting frequency is higher. This suggests that providing supervisors with more information about employee exploration by increasing reporting frequency is not necessarily sufficient to increase employee exploration.

Second, I contribute to the literature on discretionary evaluations by providing evidence consistent with an additional cost of the uncertainty inherent in discretionary evaluations. The lack of mutual understanding between supervisors and employees about how specific outcomes will be rewarded is a documented cost of discretionary evaluations (Bol, 2008; Gibbons & Henderson, 2012; Luft et al., 2016). Previous accounting literature highlights three consequences of this uncertainty. First, risk-averse employees are likely to reduce their effort because of the uncertainty related to how results will be interpreted and rewarded (Bol, 2008). Second, employees may not understand which decisions will be rewarded by the supervisors and, as a result, fail to implement those decisions (Luft et al., 2016). Third, employees may not develop accurate predictions about their evaluations and, therefore, experience negative surprises when discovering their actual evaluation (Luft et al., 2016). My results suggest a fourth cost. Employees can fail to anticipate which actions supervisors will reward in a multitasking setting. This makes supervisors less effective at directing employee effort through their bonus decision than previously thought.

Third, I contribute to the growing behavioral literature that examines the value of providing supervisors with additional information (Casas-Arce et al., 2017; Hecht et al., 2020; Luft et al., 2016). Agency theory suggests that when companies report additional information, supervisors use all this information to increase evaluation accuracy (Feltham

& Xie, 1994; Golman & Bhatia, 2012). I find that supervisors do not necessarily use all this information to reward effortful employee actions, preferring instead to partially reward uncontrollable outcomes. This suggests that providing additional information to supervisors may be less beneficial than predicted by formal models.

This study also provides relevant insights for practice. I investigate the effect on organizational outcomes of manipulating reporting frequency, a relatively cheap intervention. I find that increasing reporting frequency is not sufficient to improve employee behavior. Given that increasing reporting frequency can also deteriorate employee behavior (Cadsby et al., 2019; Hecht et al., 2021), organizations should carefully consider whether they should increase reporting frequency even if they can do so at a relatively cheap price.

Future research can build on my study in several ways. First, employee-supervisor dyads in my experiment could not develop a relational contract beyond the two evaluation decisions (Gibbons & Henderson, 2012). More evaluation periods in a similar situation may allow supervisors and employees to develop a shared understanding of how the profit information will be used in future evaluations. This could increase the usefulness of increasing reporting frequency because employees have more chances to learn how supervisors will evaluate them.

Second, my study focuses on a setting in which employees' actions produce an immediate response in the outcome measure. Examining a setting in which actions produce delayed responses and employees can learn from experience would be interesting to researchers given how difficult motivating learning in such settings is likely to be. If actions produce a delayed response, learning will become difficult for most employees (even in the absence of any incentive problem) (Gibson, 2000). Supervisors' task would also become more complex given that current period outcomes are no longer informative about the employees' actions in the current period.

Appendix 1 - Python Optimal Choice Calculation

```
from itertools import permutations
PROG= [1.1 , 1.1 , 1.1 , 1]
A = [0,0,60,60,80,80,100,100,120,120,400]
totalPeriods=10
explorationNr = 1
SOLUTION = []
MAX = max(A)
KNOWN = 200
#function for calculating the outcome of exploitation
def calculateExploit(value, timesValueAppeared, remainingTimes):
    s = 0
    for i in range(timesValueAppeared - 1, timesValueAppeared - 1 + remainingTimes, 1):
        if i > len(PROG) - 1:
            s = s + value
            continue
        s = s + value
        value = int(round(value * PROG[i]))
    return s
#Calculating the expected outcome of no exploration
SOLUTION.append([calculateExploit(KNOWN, 1, totalPeriods),0])
#creates a dictionary with the number of instances of
#every number in the grid (e.g. 400 appears once and 0 appears twice)
numbersDictionary = dict()
for nr in A:
    if nr in numbersDictionary:
        numbersDictionary[nr] = numbersDictionary[nr] + 1
    else:
        numbersDictionary[nr] = 1
while explorationNr <= totalPeriods:
    expectedOutcome = 0
    #calculates all possible draws given the number of explorations
    #(e.g. for two explorations it looks like (0,0),(0,20) etc)
    arrangements = set(list(permutations(A, explorationNr)))
```

```
    pTotal=0
    for arrangement in arrangements:
        currentNumbersDictionary = dict(numbersDictionary)
        S = 0
        P = 1
        maxNr = KNOWN
        periodsFound = 0
        aLength = len(A)
        for count, exploredNr in enumerate(arrangement, start=1):
            if periodsFound == 0:
                if exploredNr == MAX:
                    maxNr = MAX
                    periodsFound = periodsFound + 1
                S = S + exploredNr
            else:
                S = S + maxNr
                if periodsFound < len(PROG) - 1:
                    maxNr = maxNr * PROG[periodsFound - 1]
                periodsFound = periodsFound + 1
            P = float(P) * float(currentNumbersDictionary[exploredNr] )/ aLength
            currentNumbersDictionary[exploredNr] = currentNumbersDictionary[exploredNr] - 1
            aLength = aLength - 1
            currentPeriod=count
        if periodsFound == 0:
            S = S + calculateExploit(maxNr, 1, totalPeriods - currentPeriod)
        else:
            S = S + calculateExploit(maxNr, periodsFound, totalPeriods - currentPeriod)
        pTotal=float(pTotal)+float(P)
        expectedOutcome = expectedOutcome + S * P
    SOLUTION.append([expectedOutcome, explorationNr])
    explorationNr = explorationNr + 1
print(SOLUTION)
```

Chapter 3

The Effects of Reporting Structure and Reporting Frequency on Discretionary Rewards and Employee Effort

3.1 Introduction

Performance benchmarks have a substantial influence on discretionary evaluations (Chun et al., 2018; Gilbert et al., 1995; Goffin & Olson, 2011; Zell & Alicke, 2009). Yet, we know little about how performance benchmarks are affected by control system design choices. We argue that two characteristics of the organization's control system, the reporting structure (span of control) and the reporting frequency, determine which benchmarks are available, and therefore influence how employees are evaluated and rewarded. Examining the effects of span of control and reporting frequency on discretionary evaluations is timely because technological advancements have made it possible to widen supervisors' span of control (Garicano, 2000) and to report more frequently (Hecht et al., 2020).

We further argue that employees will anticipate how these control system features affect supervisors' evaluations and adjust their effort levels accordingly. A substantial body of accounting literature has examined how control system features change discretionary evaluations (Bailey et al., 2011; Bol et al., 2016; Demeré et al., 2019; Krishnan et al., 2005; Libby et al., 2004; Moers, 2005). However, less is known about whether these changes in discretionary evaluations also affect employee behavior (Arnold et al., 2018; Bol, 2011; Chan, 2018). We add to this literature by examining whether employees adjust their effort levels in *anticipation* of how control systems will affect their evaluations. Such knowledge is important because it helps us predict how changing control systems in a discretionary evaluation setting affects employee behavior in the short run.

Two important sources of benchmark levels are the performance of an employee's peers (social comparisons) and an employee's own past performance (temporal comparisons) (Chun et al., 2018; Zell & Alicke, 2009). We define span of control as the number of employees managed by a supervisor (Guo et al., 2020; Hannan et al., 2010). Span of control influences the availability of information about peer performance that can serve as a benchmark. As a result, we expect span of control affects how supervisors evaluate the performance of their employees such that, for the same performance, an employee will receive a different rating depending on whether his supervisor presides over few or many other employees. We define reporting frequency as the number and the granularity of reports generated about an employee's performance in a given time interval. Reporting frequency influences the availability of information about an employee's past performance that can serve as a benchmark. When reporting frequency increases, supervisors can observe that an employee's performance is lower in some periods and higher in others. We expect supervisors to pay more attention and put more weight on the periods in which the employee did relatively poorly because supervisors are influenced by negativity bias in their evaluations (Baumeister et al., 2001; Kaplan et al., 2012, 2018). As a result, we predict that employees will receive lower evaluations when reporting frequency increases. Our final prediction about supervisors' evaluations involves an interaction between span of control and reporting frequency. We argue that, as span of control widens, supervisors will pay less attention to employees' performance across time because supervisors will find it less cognitively demanding to focus on a single type of benchmark (Chase & Simon, 1973; Shanteau, 1988) and because they find social comparisons more relevant than temporal comparisons (Zell & Alicke, 2009). Thus, we expect that increasing reporting frequency lowers evaluations less when span of control is wider.

We further theorize that employees will anticipate how these control choices will affect their evaluations and choose their effort levels to maximize their expected payoff. Specifically, we predict that (1) a wider span of control will cause employees to increase their effort, (2) a higher reporting frequency will cause employees to choose effort levels that reduce the variation in their performance, and (3) a higher reporting frequency will reduce variation in performance less when span of control is wider.

We test our hypotheses using two experiments. In an online case-based experiment, we examine how participants evaluate hypothetical employees differently depending on our control system manipulations. We find support for our prediction that discretionary reward allocations are affected by the supervisors' span of control. Specifically, strong performers receive higher rewards and weak performers receive lower rewards from supervisors with wider spans of control. In contrast to our theory, reporting frequency does not affect reward allocations and the effect of reporting frequency is not moderated by span of control. In a laboratory experiment, we examine employee effort under different control systems. Employees perform a real-effort task and are evaluated by supervisors. The results from this experiment suggest that, in defiance of our theory, employee effort is not affected by their supervisor's span of control or the frequency of performance reports.

We believe that despite the lack of support for many of our hypotheses our study provides several important insights for management accounting research and practice. First, we contribute to the discretionary evaluation literature by investigating how the reporting structure affects discretionary performance evaluations. Specifically, we find that when companies decide to 'flatten their organization' (Hannan et al., 2010), supervisors give higher (lower) evaluations to their best (worst) performing employees. The human-resources literature finds mixed results when examining the effect of span of control on evaluations (Ellington & Wilson, 2017; Judge & Ferris, 1993; Lahuis & Avis, 2007; O'Neill et al., 2012). We help explain these mixed results by providing evidence that span of control affects evaluations differently depending on an employee's standing among their peers.

Second, we contribute to the discretionary evaluation literature by examining whether employees adjust their effort levels in anticipation of how different control systems will affect supervisors' evaluations. In the case-based experiment, we find that supervisors change their evaluation patterns when span of control widens. If employees accurately anticipate how supervisors will evaluate them, as it is sometimes assumed in the literature (e.g. Baiman & Rajan, 1995), we should have observed an effect of span of control on employee effort in our lab study. Our results show no evidence of such an effect however. This emphasizes the importance of developing a more comprehensive theory about how changes in discretionary evaluations affect employee behavior.

Third, we contribute to the emerging literature on the effects of changing reporting frequency within the firm. While reporting frequency at the firm level has been examined in a financial accounting context (Ernstberger et al., 2017; Kajüter et al., 2019; Wagenhofer, 2014), our study is one of the first to examine reporting frequency within the firm. Our findings complement the conclusions of a recent study on the subject. Hecht et al. (2020) examine the motivation effects of increasing reporting frequency when performance depends primarily on employee ability. We examine how reporting frequency affects employee motivation when performance depends primarily on employee effort. While Hecht et al. (2020) find that a higher reporting frequency decreases employee motivation, we find no effect of reporting frequency on employee effort. Thus, our results expand our understanding of how reporting frequency affects employee motivation and suggest that this effect may depend on the type of task, specifically on the relative importance of effort and ability as antecedents of employee performance.

3.2 Background and Hypothesis Development

3.2.1 Background

Supervisors often have some discretion over the rewards received by their employees. For example, they can determine employees' annual bonuses, salary increases, or eligibility for promotions (Baiman & Rajan, 1995; Baker et al., 1994; Gibbs et al., 2004; Hecht et al., 2021). Organizations provide supervisors with this discretion because it allows supervisors to incorporate non-contractible information in employee compensation (Bol, 2008) and because explicit contracts are often prohibitively costly to write (Choi et al., 2016).

Discretionary evaluation of employee performance inherently entails the use of benchmarks. Existing research suggests two important sources of benchmark levels of performance are the performance of the employee's peers and the employee's own past performance. In other words, when evaluating the achievements of an employee, supervisors tend to make two types of comparisons: (1) social comparisons where they compare how well an employee is performing in relation to their peers, and (2) temporal comparisons where supervisors examine how an employee's performance changes over time (Chun et al., 2018; Zell & Alicke, 2009). Whether such benchmark information is available and therefore has the potential to influence supervisors' evaluations, depends on an organization's control system. First, the availability of information about peers will depend on the span of control of a supervisor. A supervisor who needs to evaluate a relatively large number of employees will use peer comparison differently than a supervisor who only needs to assess the performance of a few employees. Second, the availability of information for temporal comparisons depends on the performance reporting frequency in an organization. More frequent reporting provides supervisors with more signals of employee performance and allows them to identify trends and fluctuations in an employee's performance over time.[1]

3.2.2 Supervisors' Bonus Allocations

Span of Control

Span of control refers to the number of employees that report directly to the same supervisor (Guo et al., 2020; Hannan et al., 2010). A supervisor with a wider span of control needs to pay attention to the performance of a larger number of employees than a supervisor with a narrower span of control. Therefore, when evaluating employees, supervisors with wider spans of control have access to more information about how other employees in the same position are performing.[2]

Span of control could affect overall evaluations in at least two ways. First, a wider span of control could cause supervisors to become more lenient. Supervisors with wider spans of control evaluate more employees with below-average levels of performance. Therefore, supervisors with wider spans of control need to give more below-average evaluations if they want to solely evaluate employees based on performance. For example, a non-lenient supervisor who evaluates ten employees needs to give below-average bonuses to a maximum of five employees while a non-lenient supervisor who evaluates four employees needs to give a below-average bonus to a maximum of two employees. Given that employees confront supervisors more when they receive below-average evaluations (Bol et al., 2016; Sebald &

[1]Traditional agency models suggest that supervisors only perform social comparisons to filter out the effects of common, uncontrollable events from employees' evaluations (e.g. Holmstrom, 1982). More recent economics and psychology research suggests that people use social comparisons for more than filtering out noise from their evaluations (Bhargava & Fisman, 2014; Gilbert et al., 1995; Goffin & Olson, 2011).

[2]Supervisors with a narrow span of control could, in principle, collect the same amount of information as supervisors with wide spans of control by observing the performance and circumstances of employees whom they do not need to evaluate. However, collecting such information would be time-consuming for supervisors (Maas et al., 2012) and would likely decrease performance on their other duties (Hofmann & Indjejikian, 2018). Moreover, supervisors likely do not always pay attention to all the useful information available to them (Berger, 2019; Zureich, 2020). As a result, supervisors may not realize that paying attention to the performance of employees from other teams may improve their evaluations. Therefore, consistent with prior studies (Hannan et al., 2010; O'Neill et al., 2012), we expect supervisors with wider spans of control observe more information than supervisors with narrow spans of control.

Walzl, 2014), supervisors with wider spans of control will anticipate that they would be confronted more often if they would reward employees based solely on their performance. Supervisors could choose to become more lenient when their span of control widens to avoid such confrontations.

Second, a wider span of control could cause supervisors to base evaluations more on employee performance. As mentioned earlier, supervisors with wider spans of control have access to more information about how other employees in the same position are performing. Therefore, comparing employees will be more informative for supervisors with wider spans of control (O'Neill et al., 2012) which, in turn, could increase evaluation accuracy (Golman & Bhatia, 2012).

Consistent with the idea that the effect of span of control on evaluations is complex, previous literature finds mixed results when examining how span of control affects evaluations. Gong et al. (2019) find that supervisors with wider spans of control are more lenient. O'Neill et al. (2012) find that supervisors with wider spans of control give evaluations that are more in line with employee performance. Other studies find no effect of span of control on evaluations (Ellington & Wilson, 2017; Judge & Ferris, 1993; Lahuis & Avis, 2007).

To help explain these mixed results, we propose that the effect of span of control on evaluations depends on the relative standing of the employee in the performance distribution. We argue that relatively high performing employees will receive more favorable evaluations, and thus higher discretionary rewards, as the span of control of their supervisor widens. On the contrary, we expect relatively low performing employees to receive less favorable evaluations and lower discretionary rewards if their supervisor has a wider span of control.

When evaluating employees, supervisors generally try to be accurate and fair (Chan, 2018; Golman & Bhatia, 2012; Maas et al., 2012). They are particularly concerned about relative fairness, i.e., they want to provide evaluations and rewards that accurately reflect employees' performance compared to their peers (Bol et al., 2015, 2016; Maas et al., 2012).

Because of this concern for relative fairness, supervisors will likely avoid allocating exceptionally high or low bonuses unless they have sufficient evidence that employees' performance is exceptional. This is because, in the presence of uncertainty about how employees' performance levels should be assessed, supervisors will prefer to offer relatively high and similar evaluations to all employees (Bol et al., 2016; Golman & Bhatia, 2012).

For a supervisor of a small team, it will be difficult to assess whether the high and low performing employees in their team are truly exceptional. For example, suppose a supervisor evaluates two employees. While one employee may perform substantially better than the other, it is possible that if the team were larger it would become apparent that both employees are actually average performers, or that both are top performers or weak performers. As the supervisor's span of control increases, it will become clearer whether the best and worst performers are exceptional. Supervisors will therefore become more comfortable with allocating exceptionally high or low bonuses to these employees. Thus, as the supervisor's span of control increases, top performers should expect higher ratings and weak performers should expect lower ratings. This reasoning leads to our first two hypotheses:

H1a: Best performing employees receive higher discretionary rewards from supervisors with wider spans of control.

H1b: Weakest performing employees receive lower discretionary rewards from supervisors with wider spans of control.

Reporting Frequency

We define reporting frequency as the number and the granularity of reports generated about an employee's performance in a given time interval.[3] For example, assuming an evaluation period of a year, the lowest frequency would be a single report about the employee's performance at the end of the year. At the other – high frequency - extreme, the employee's performance could be measured and reported on a weekly, daily, or even real-time basis.[4,5]

The higher the reporting frequency, the more signals about an employee's performance a supervisor receives. Increasing reporting frequency thus facilitates temporal within-employee comparisons because it gives supervisors more information about the variability across time in employees' performance. Increasing reporting frequency will likely result in supervisors observing more periods with both above and below average performance. This is partly because factors such as fatigue cause employees' effort to vary across time (Collewet & Sauermann, 2017; Fritz et al., 2013; Pencavel, 2015) and because some of these factors offset each other when performance is reported less frequently (Arya & Glover, 2014).

Of course, supervisors could just ignore the intermediate reports and focus only on the aggregate performance level when evaluating an employee's performance. Note that in this case, the supervisor would implicitly give equal weight to each intermediate performance signal. However, we predict that supervisors will not ignore the intermediate reports but

[3] Reporting frequency is different from feedback frequency. Feedback frequency refers to the frequency with which an employee receives feedback about his or her actions. A large literature in accounting and management examines how feedback frequency affects employee performance (Anand et al., 2019; Casas-Arce et al., 2017; Lam et al., 2011; Lurie & Swaminathan, 2009). We do not focus on feedback frequency in this study and keep it constant across the experimental conditions.

[4] By investigating the effects of increasing reporting frequency, we examine the cumulative effect of differences in two reporting dimensions: information aggregation and the time interval between reports. This is an intentional choice because companies likely change both these dimensions simultaneously. More frequent reports naturally involve less information aggregation. Less-aggregate reports likely also involve the possibility that the supervisor will observe employee performance more frequently. For example, most supervisors likely do not need to wait until the end of the month to learn about a salesperson's performance on a given day if the company collects daily information about salespeople's performance.

[5] Similar to Hecht et al. (2020), we examine a setting where employees cannot misreport their performance. Instead, the reporting system automatically captures and reports the employees' performance.

will instead compare employees' performance levels over time, and place disproportionate weight on performance signals that are below average, or indicate a downward trend. Existing research in psychology and accounting shows that people tend to give more weight to negative information than to equal-in-magnitude positive information. This tendency, which has sometimes been labeled the negativity bias (e.g. Amabile & Glazebrook, 1982; Skowronski & Carlston, 1989), has been observed in many different settings (Baumeister et al., 2001), including discretionary performance evaluation settings (Kaplan et al., 2012, 2018). Baumeister et al. (2001) argue that negativity bias is likely hard-wired in people because an asymmetrical response to opportunities and threats makes sense from an evolutionary perspective: "A person who ignores the possibility of a positive outcome may later experience significant regret at having missed an opportunity for pleasure or advancement, but nothing directly terrible is likely to result. In contrast, a person who ignores danger (the possibility of a bad outcome) even once may end up maimed or dead" (Baumeister et al., 2001, p. 325).

If supervisors weigh below-average performance signals more heavily than above-average performance signals, we should expect the overall evaluation, and therefore the discretionary reward for an employee to be lower as reporting frequency increases. A very similar argument is made by Bentley & Stubbs (2020) who note that if supervisors are loss averse (Tversky & Kahneman, 1981), the higher salience of short-window losses under more frequent reporting could lead to lower discretionary evaluations. As we have no reason to expect the effect of reporting frequency to be different for top performers and weaker performers, we test the following hypothesis:

H2: Increasing reporting frequency decreases the discretionary rewards that supervisors award to employees.

Reporting Frequency and Span of Control

Finally, we consider the joint effect of span of control and reporting frequency on evaluations and discretionary rewards. When span of control is narrow and reporting frequency is low, supervisors have few benchmarks to compare an employee's performance to. Both span of control and reporting frequency increase the number of performance benchmarks. We believe that an increase in cross-sectional comparison opportunities due to a wider span of control, will be more salient when reporting frequency is low. Similarly, an increase in temporal comparison opportunities due to increased reporting frequency will be more salient when a supervisor has a narrow span of control.

This negative interaction effect is supported by cognitive psychology research that suggests that cognitive limitations lead people who need to process a large number of information cues to use a 'divide and conquer' strategy (e.g. Chase & Simon, 1973; Shanteau, 1988). This means they use simplifying heuristics and deliberately focus on certain pieces of information while ignoring others. For example, a supervisor with a wide span of control may choose to deliberately ignore the intermediate performance reports and just compare her many subordinates based on their overall performance. Alternatively, she may choose to focus on individual performance trends over time, deliberately refraining from making explicit cross-sectional comparisons.

The only study that we are aware of that has explicitly examined evaluations in the presence of both social and temporal comparison opportunities is Zell & Alicke (2009). This study found that evaluators tended to ignore temporal comparisons, instead focusing exclusively on cross-sectional comparisons of performance levels. This suggests that if an employee stands out among his peers (either positively or negatively), supervisors will pay less attention to how that employee's performance developed over time. Thus, the discretionary reward of employees who clearly are top performers or weak performers, are less likely to be adjusted downward as reporting frequency increases. In summary, we expect the negative effect of reporting frequency to be weaker if span of control is wider. This leads to the following hypothesis:

H3: Increasing reporting frequency has a more negative effect on discretionary rewards when span of control is narrower.

3.2.3 Employee Effort

Span of Control

Our first set of hypotheses predicts how span of control and reporting frequency affect supervisors' discretionary evaluations. We now turn to the effects of these factors on the behavior of employees, specifically on employee effort provision. Underlying the following predictions is the premise that employees will anticipate supervisors' discretionary reward decisions, and choose their effort levels to maximize their expected payoff.

First, we expect employees to put in more effort if their supervisor has a wider span of control. Our reasoning above suggests that, as span of control widens, supervisors give higher discretionary rewards to their best performers and lower rewards to their weaker performers. Employees of supervisors with a wider span of control thus have two reasons to provide relatively high levels of effort: they need to prevent being the weakest performer and face additional benefits from being the strongest performer. This reasoning leads to the following hypothesis:

H4: Employees increase their effort when they are evaluated by supervisors with a wider span of control.

Reporting Frequency

Second, our reasoning above suggests that reporting frequency leads to lower rewards because supervisors weigh below-average performance signals more heavily than above-average performance signals. If employees anticipate that supervisors will place disproportionate weight on clearly below-average performance signals, they will try to avoid such signals. Because with more frequent reporting it is more likely that a temporary drop in performance is observable to the supervisor, employees will try harder to avoid such drops in performance as reporting frequency increases. For example, an employee who knows that his performance is measured and reported to his supervisor at the end of each week will adjust his effort such that sudden drops in performance are prevented in any week. Compare this to an employee whose performance is only reported every month. This employee can 'afford' to slack off a little in certain weeks. While this may have a marginal effect on his monthly performance, this drop will be much less salient to a supervisor when performance is reported monthly than when it is reported weekly.

Notably, avoiding temporary drops in performance may - intentionally or unintentionally – also lead to fewer and lower peaks in performance. First, performance peaks make performance valleys seem deeper and therefore employees will try to avoid them. Second, peaks may cause valleys as employees need some time to recover from providing exceptionally high effort levels. Therefore, we expect employees will choose their effort levels such that they deliver more consistent results (lower variation in performance over time) when reporting frequency is higher.

H5: Increasing reporting frequency decreases variation in employee effort over time.

In a recent paper, Hecht et al. (2020) also investigate the effect of reporting frequency on employee performance and find that a higher frequency decreases employee performance. The authors argue that potential losses (periods with low performance) become more salient to employees when reporting frequency increases because they are closer and therefore more painful. As a result, employees will be more likely to focus on avoiding losses. This 'avoidance orientation' leads to low task absorption and anxiety which, in turn, lead to lower performance. Importantly, reporting frequency only decreases performance when employees know that reports will be used to evaluate their task-related skill, i.e., their ability.

The setting in Hecht et al. (2020) is different from ours as in their paper performance is mostly determined by employee ability while we focus on a setting where performance is mostly determined by effort. We do not predict that a higher reporting frequency will decrease employee performance because we do not expect that the negative effects documented by Hecht et al. (2020) will materialize in our setting. When performance is mostly determined by effort, we expect reporting losses to be less painful to employees as these losses do not reflect shortcomings in their ability. Therefore, while we expect that employees will try to avoid reporting losses to increase their bonuses, we do not expect that trying to do so will cause them to experience performance-decreasing anxiety.

Reporting Frequency and Span of Control

Our third and final prediction is that there is an interaction effect of span of control and reporting frequency on the variation in (but not the level of) employee effort. H3 states that supervisors' reward decisions will be more strongly affected by reporting frequency when span of control is narrower. We predict that employees will anticipate that their fluctuations in performance will receive less attention from supervisors as span of control widens and adjust their effort level accordingly. Thus, an employee whose superior has a narrower span of control will feel a more urgent need to reduce variation in his performance over time than an employee whose supervisor has a wider span of control. This is reflected in our final hypothesis:

H6: Increasing reporting frequency has a more negative effect on variation in employees' effort overtime when span of control is narrower.

It is important to note that our hypotheses about employee effort are not without tension. As indicated, a core assumption underlying these hypotheses is that employees will anticipate (either correctly or not) that supervisors' reward allocations will be affected by elements of the control system (i.e., span of control and reporting frequency). A considerable body of literature in psychology and economics shows that anticipation of other individuals' decisions and action strategies is not straightforward and that there is much variation in the extent to which individuals try to understand a situation from another person's perspective (e.g. Camerer et al., 2015; Cardinaels et al., 2018; Zhang et al., 2012). Moreover, this literature shows that failure to adequately anticipate others' behavior is typically not fully mitigated by economic incentives (Camerer et al., 2015). If employees fail to put themselves in the shoes of their supervisor, then span of control and reporting frequency may not affect their effort levels.

We test our hypotheses using two experiments. Specifically, H1, H2, and H3 are tested in an online experiment with participants recruited on the Prolific platform, and H4, H5, and H6 are tested in an interactive lab experiment in which students engage in a real effort task. We discuss each experiment and its results separately below.

3.3 Experiment One

3.3.1 Design

We test the hypotheses about supervisors' discretionary reward allocations (H1, H2, and H3) using a case-based experiment with a 2 (span of control: narrow or wide) × 2 (reporting frequency: low or high) between-subjects design. Participants assumed the role of a regional manager of a retail company. Their task was to assess the performance of a group of store managers and to allocate a bonus to each of them. We manipulated span of control and reporting frequency in the case scenario. In the narrow (wide) span of control condition, participants allocated a separate bonus to two (five) store managers. In the low reporting frequency condition, participants received net income reports that contained information about the aggregate net income for the entire year. In the high reporting frequency condition, participants received net income reports that contained information about net income for every quarters. The dependent variables are the bonuses assigned to the lowest and highest performer, as explained in more detail below. The design and procedures were reviewed and accepted by the research ethics committee (IRB) of the University of Amsterdam.

Participants and Procedures

We recruited participants on the online platform Prolific. Using Prolific's pre-screening feature (Palan & Schitter, 2018), we limited our target sample to workers who reported having the authority to give instructions to at least one employee at work. We also restricted our sample to participants residing in the United States or the United Kingdom and using a desktop or laptop computer (not a mobile device). In total, 334 Prolific qualified workers started the study. After reading the instructions, participants had to answer a few understanding check questions on which they needed to score 100% to move to the main task. Ten workers quit and did not attempt to answer the understanding questions. 94 participants failed to correctly answer all the understanding questions on their first attempt and were requested to read the instructions another time. Out of these 94 participants, four did not attempt to answer the questions again and 28 failed to correctly answer all attention questions on their second attempt and were not allowed to continue. One participant was removed from the sample because of technical issues.[6] This leaves our final sample size at 291 participants. Of these, 154 (52.9%) are male and 137 (47.1%) are female. On average, participants are 39.85 years old and report an average of 19.56 years of work experience and 9.18 years of supervisory experience. The average number of supervised employees as work is 15.88.

Prospective participants who accepted the task on Prolific were directed to a Qualtrics survey and the Qualtrics software randomly assigned them to one of the four conditions.[7] Participants first provided their informed consent, and then read the study's basic instruc-

[6]This participant restarted the experiment after completing the understanding checks likely because of a technical issue. Because he saw a different manipulation on his second attempt at completing the study, we decided to exclude this participant from our sample.

[7]Participants' characteristics do not differ between treatments with respect to age, gender, work experience, employment status (full-time worker or part-time worker), student status, supervisory experience, number of subordinates, country of residence (the United Kingdom or the United States) their first language (English or other). Results of a multiple linear regression do not reveal an association between the measured characteristics and assignment to one of the reporting frequency conditions (F(9, 281) = 1.09, p = 0.372, R^2 = 0.03) or one of the span of control conditions (F(9, 281) = 0.91, p = 0.521, R^2 = 0.02).

tions and the case scenario. They then had two attempts to score 100% on a quiz consisting of six understanding check questions. Participants who passed the understanding check subsequently filled out a questionnaire that contained their responses to the case scenario (our dependent variables) as well as questions about their decision-making process and demographic questions. Participants who completed the study received £1 for an average of approximately 9.5 minutes of their time.

Task and Manipulations

Participants were asked to assume the role of a regional manager at JSM Inc., a hypothetical retail company. JSM Inc. was described as operating a chain of 150 stores that sell apparel, accessories, and shoes in different regions of the country. In their role as regional manager, participants needed to review the performance - and determine the annual bonus - of the store managers in their region. They had full discretion in allocating a bonus between \$0 and \$10,000 to each store manager. They were asked to base their bonus on the financial performance of each store and were provided with a performance report with financial performance information for each store.[8]

We manipulated the independent variables by varying the number of stores and the specifics of the performance report. Participants in the narrow span of control condition evaluated two store managers while participants in the wide span of control condition evaluated five store managers. In both span of control conditions, participants learned that the best performing store in their region was Store A (annual net income of \$425,000) and the worst performing store was Store B (annual net income of \$280,000). Participants in the

[8] Participants were told to assume that all stores were of a similar size and faced similar economic and market conditions. They were also asked to imagine that their own payoff depended on the average performance of the store managers. To ensure that participants were not motivated to deliberately underestimate store manager performance in order to use the unallocated bonus for their own interest, they were informed that money that is not rewarded to the store managers is redistributed to the company bonus pool. Finally, to ensure that our results are not driven by participants' anticipation that store managers will compare their bonuses (Bol et al., 2015, 2016), the case informed participants that JSM had a policy to keep all bonus decisions strictly confidential.

wide span of control condition also evaluated and therefore learned about the performance of the managers of Stores C, D, and E who generated net incomes of \$395,000, \$335,000, and \$365,000 respectively. In the low reporting frequency condition, participants received aggregate annual net income reports for each store. In the high reporting frequency condition, they also received information about the net income of each store in each specific quarter. Naturally, each store had at least one quarter in which the performance was lower than their own average performance during the year. Specifically, Store A (the strongest performer) and B (the weakest performer) each performed clearly below average in one quarter, whereas Stores C, D, and E each performed below average in two quarters. The dips in performance for different stores were not concentrated in specific quarters to avoid the conclusion that dips in performance were due to seasonality. Table 3.1 presents the performance reports as they were seen by the participants.

3.3.2 Results

Table 3.2 presents descriptive statistics on the bonus allocated to the store managers in each of the four conditions. From Table 3.2 it is clear that in each condition the bonus allocated to the manager of Store A (the best performing store) is higher than the bonus allocated to Store B (the worst performing store). In fact, t-tests reveal that the difference between the bonus assigned to the best performer and the bonus assigned to the worst performer is significantly different from zero in all four conditions (all $p < 0.001$).

We test our hypotheses using factorial ANOVAs with the two manipulations as factors and the bonuses assigned to the managers of Store A (*BonusA*) and Store B (*BonusB*) and the combined bonus of the managers of stores A and B (*AggregateBonus*) as dependent variables. We create two dummy variables: *WideSpan* and *HighFrequency*. The dummy variable *WideSpan* takes on value 1 if span of control is wide (five employees) and value 0 if span of control is narrow (two employees). The dummy variable *HighFrequency* takes on the value 1 if the observation is from a high reporting frequency condition (four quarterly

Table 3.1: The Performance Report Seen by Participants in Experiment One

Wide Span of Control, High Reporting Frequency Condition

	Store A	Store B	Store C	Store D	Store E
Net Income in Quarter 1 (JAN - MAR)	$120,000	$75,000	$80,000	$110,000	$125,000
Net Income in Quarter 2 (APR - JUN)	$110,000	$40,000	$95,000	$75,000	$85,000
Net Income in Quarter 3 (JUL - SEP)	$80,000	$85,000	$120,000	$105,000	$80,000
Net Income in Quarter 4 (OCT - DEC)	$115,000	$80,000	$100,000	$45,000	$75,000
Net Income for the Year	$425,000	$280,000	$395,000	$335,000	$365,000

Wide Span of Control, Low Reporting Frequency Condition

	Store A	Store B	Store C	Store D	Store E
Net Income for the Year	$425,000	$280,000	$395,000	$335,000	$365,000

Narrow Span of Control, High Reporting Frequency Condition

	Store A	Store B
Net Income in Quarter 1 (JAN - MAR)	$120,000	$75,000
Net Income in Quarter 2 (APR - JUN)	$110,000	$40,000
Net Income in Quarter 3 (JUL - SEP)	$80,000	$85,000
Net Income in Quarter 4 (OCT - DEC)	$115,000	$80,000
Net Income for the Year	$425,000	$280,000

Narrow Span of Control, Low Reporting Frequency Condition

	Store A	Store B
Net Income for the Year	$425,000	$280,000

This tables present the performance reports as they were seen by the participants in Experiment One. In the wide span of control condition, participants observed the performance of all five stores while in the narrow span of control condition, participants only observed the performance of stores A and B. In the high reporting frequency condition, participants observed the performance for each quarter and the aggregate net income for the year while in the low reporting frequency condition, participants only observed the aggregate net income for the year.

Table 3.2: Descriptive Statistic: Bonuses in Experiment One

Condition	Low Frequency			High Frequency			Overall		
Narrow Span of Control		n	73		n	74		n	147
	Store A	mean	7,034	Store A	mean	7,464	Store A	mean	7,250
		sd	2,226		sd	2,135		sd	2,184
	Store B	mean	4,130	Store B	mean	4,360	Store B	mean	4,246
		sd	1,935		sd	1,859		sd	1,894
	Aggregate	mean	11,164	Aggregate	mean	11,824	Aggregate	mean	11,496
		sd	3,906		sd	3,200		sd	3,571
Wide Span of Control		n	73		n	71		n	144
	Store A	mean	8,799	Store A	mean	8,187	Store A	mean	8,497
		sd	1,883		sd	2,329		sd	2,129
	Store B	mean	3,560	Store B	mean	3,934	Store B	mean	3,744
		sd	2,434		sd	2,672		sd	2,552
	Aggregate	mean	12,359	Aggregate	mean	12,121	Aggregate	mean	12,241
		sd	3,496		sd	3,914		sd	3,697
Overall		n	146		n	145		n	291
	Store A	mean	7,916	Store A	mean	7,818	Store A	mean	7,867
		sd	2,237		sd	2,253		sd	2,242
	Store B	mean	3,845	Store B	mean	4,151	Store B	mean	3,998
		sd	2,210		sd	2,295		sd	2,254
	Aggregate	mean	11,761	Aggregate	mean	11,969	Aggregate	mean	11,865
		sd	3,742		sd	3,558		sd	3,647

This table presents the descriptive statistics of Experiment One. It reports descriptive statistics on the aggregate and individual bonuses allocated to the best and worst performing store managers across the two span of control conditions and across the two reporting frequency conditions.

performance reports) and 0 if the observation is from the low reporting frequency condition (one annual performance report). The results are presented in Tables 3.3, 3.4 and 3.5.

H1a predicts that best performing employees will receive higher discretionary rewards from supervisors with wider spans of control. Consistent with this hypothesis Panel A of Table 3.3 shows that there is a main effect of *WideSpan* on *BonusA* ($F = 24.38$, $p < 0.001$). As is clear from Table 3.2, the average bonus assigned to the top performer increases from 7,250 (sd = 2,184) to 8,497 (sd = 2,129) when span of control widens. Next, H1b predicts that the weakest performing employees will receive lower discretionary rewards from supervisors with wider spans of control. Consistent with this hypothesis, Panel A of Table 3.4 shows that there is a marginally significant main effect of *WideSpan* on *BonusB* ($F = 3.58$, $p = 0.06$). Table 3.2 shows that on average the bonus of the weakest performing store manager was 4,246 (sd = 1,894) when span of control was narrow and 3,744 (sd = 2,552) when span of control was wide. Together, these results provide support for H1a and H1b.

H2 predicts that increasing reporting frequency decreases the discretionary rewards that supervisors award to employees. To test this hypothesis we first look the aggregate of the bonuses assigned to the managers of stores A and B.[9] From Table 3.2 it is clear that, inconsistent with this hypothesis, the average aggregate bonus in the low frequency condition is 11,761 (sd = 3,742), while in the high frequency condition it is 11,969 (sd = 3,558). To formally test the hypothesis we run an ANOVA with *AggregateBonus* as the dependent variable. The results are in Table 3.5. This table shows that the main effect of *HighFrequency* on *AggregateBonus* is not significant ($F = 0.25$, $p = 0.621$). If we look at *BonusA* and *BonusB* separately, we see a similar pattern. The main effect of *HighFrequency* is insignificant in both Table 3.3 ($F = 0.13$, $p = 0.720$) and Table 3.4 ($F = 1.31$, $p = 0.253$).

[9] For our main test, we do not examine the effect of reporting frequency on the bonuses awarded to the managers of stores C, D, and E because we only observe these bonuses in the Wide Span of Control condition. An analysis on bonuses awarded to all store managers would have captured the average effect of increasing reporting frequency while span of control is wide.

Table 3.3: The Effects on the Bonus of the Best Performing Manager

Panel A - Analysis of Variance on the Bonus Allocated to the Manager of Store A

Source	Partial SS	df	MS	F	p
Model	133,300,000	3	44,443,493	9.63	<0.001
WideSpan	112,500,000	1	112,500,000	24.38	<0.001
HighFrequency	595,393	1	595,393	0.13	0.720
WideSpan * HighFrequency	19,761,996	1	19,761,996	4.28	0.039
Error	1,325,000,000	287	4,615,058		
Total	1,458,000,000	290	5,027,077		

Panel B - Simple Effects on the Bonus Allocated to the Manager of Store A

	Prediction	Contrast	df	t	p one-tailed
WideSpan when RF = Low	>0	1,765.07	144	5.172	<0.001
WideSpan when RF = High	>0	722.56	143	1.949	0.027
HighFrequency when SoC = Narrow	<0	430.78	145	1.198	0.884
HighFrequency when SoC = Wide	<0	-611.73	142	1.736	0.042

Panel A of this table presents the results of a factorial ANOVA with the two manipulations as factors and the bonus assigned to the manager of Store A, the best performing manager, as a dependent variable. Panel B of this table presents the simple effects of our manipulations.

Together, these results indicate that H2 is not supported: increasing reporting frequency does not lead to overall lower bonus assignments.

H3 predicts that there is an interaction effect of span of control on reporting frequency, such that the effects of either variable will be weaker for higher values of the other variable. We specifically expected that increasing reporting frequency would have a stronger negative effect on discretionary rewards when span of control is narrower. Note that while there is no main effect of reporting frequency, it could still be the case that reporting frequency does reduce bonus assignments in the narrow span condition only. We again first look at the aggregate of the bonuses assigned to the strongest and weakest performer. Panel B of Table 3.5 shows that the effect of reporting frequency on *AggregateBonus* is insignificant in both span of control conditions (narrow span: $t = 1.122$, one-tailed $p = 0.868$; wide span: t

Table 3.4: The Effects on the Bonus of the Worst Performing Manager

Panel A - Analysis of Variance on the Bonus Allocated to the Manager of Store B

Source	Partial SS	df	MS	F	p
Model	25,255,361	3	8,418,454	1.67	0.174
WideSpan	18,038,945	1	18,038,945	3.58	0.060
HighFrequency	6,622,456	1	6,622,456	1.31	0.253
WideSpan * HighFrequency	378,379	1	378,379	0.07	0.784
Error	1,448,000,000	287	5,045,828		
Total	1,473,000,000	290	5,080,717		

Panel B - Simple Effects on the Bonus Allocated to the Manager of Store B

	Prediction	Contrast	df	t	p one-tailed
WideSpan when RF = Low	<0	-570.14	1	1.566	0.060
WideSpan when RF = High	<0	-425.88	1	1.118	0.133
HighFrequency when SoC = Narrow	<0	229.62	1	0.734	0.768
HighFrequency when SoC = Wide	<0	373.87	1	0.878	0.809

Panel A of this table presents the results of a factorial ANOVA with the two manipulations as factors and the bonus assigned to the manager of Store B, the worst performing manager, as a dependent variable. Panel B of this table presents the simple effects of our manipulations.

= 0.385, one-tailed p = 0.351). In addition, Panel A of this table shows that the interaction effect of *WideSpan* and *HighFrequency* on *AggregateBonus* is insignificant (F = 1.11, p = 0.293). These findings provide no support for H3.

Next, we examine the interaction effect of span of control and reporting frequency for the strongest and the weakest performers separately. First, looking at the results for the top performer in Panel A of Table 3.3, we do see a significant interaction effect of *WideSpan* and *HighFrequency* (F = 4.28, p = 0.039). Looking at the simple effects reported in Panel B of this table, it is clear that the pattern of the interaction is inconsistent with our prediction in H3. While the prediction was that reporting frequency would matter more for supervisors with a narrower span of control, the results show that increasing reporting frequency only significantly reduced the bonus assigned to the stronger performer in the wide span condition

Table 3.5: The Effects on the Aggregate Bonus

Panel A - Analysis of Variance on the Aggregate Bonus Allocated to the Manager of Store A and B

Source	Partial SS	df	MS	F	p
Model	58,459,771	3	19,486,590	1.47	0.222
WideSpan	40,456,290	1	40,456,290	3.06	0.082
HighFrequency	3,246,469	1	3,246,469	0.25	0.621
WideSpan * HighFrequency	14,671,362	1	14,671,362	1.11	0.293
Error	3,799,000,000	287	13,235,373		
Total	3,857,000,000	290	13,300,041		

Panel B - Simple Effects on the Aggregate Bonus Allocated to the Manager of Store A and B

	Prediction	Contrast	df	t	p one-tailed
HighFrequency when SoC = Narrow	<0	660.40	1	1.122	0.868
HighFrequency when SoC = Wide	<0	-237.86	1	0.385	0.351

Panel A of this table presents the results of a factorial ANOVA with the two manipulations as factors and the aggregate bonus assigned to the manager of Store A, the best performing manager, and Store B, the worst performing manager, as a dependent variable. Panel B of this table presents the simple effects of our manipulations.

(t = 1.736, one-tailed p = 0.042). In the narrow span condition, the bonus assignment did not vary between low and high frequency reporting (t = 1.198, one-tailed t = 0.884). Finally, the results in Table 3.4 indicate that span of control and reporting frequency also did not interact to affect the bonus assigned to the weakest performer. The interaction effect is insignificant (F = 0.07, p = 0.784), and so are the effects of reporting frequency in both the narrow span (t = 0.734, one-tailed p = 0.768) and wide span (t = 0.878, p = 0.809) conditions. In summary, we do not find support for H3.

Additional Analyses

Our main results provide support for H1a and H1b but not for H2 and H3. Thus, we find that widening supervisors' span of control tends to increase the discretionary rewards for top performers and decrease the discretionary rewards for weaker performers. Increasing reporting frequency does not lead to lower discretionary rewards, except for top performers whose supervisor has a wide span of control. We next report some additional analyses related to our finding that span of control affects bonus assignments. The results of these analyses further substantiate the support for H1a and H1b.

First, we argued that supervisors are more likely to perceive the worst performer as an exceptionally low performer and the best performer as an exceptionally high performer when span of control widens. If this is true, we should observe that supervisors are more likely to allocate an exceptionally high (low) bonus to the best (worst) performing employee when span of control widens. Untabulated results suggest that this is indeed the case. More supervisors allocate the maximum bonus to the best performing employee in the wide span condition (74 participants, 51.39%) than in the low span condition (30 participants, 20.41%). This difference is significant (χ^2 = 30.201, p < 0.001). Similarly, significantly (χ^2 = 11.068, p = 0.001) more supervisors allocate zero bonus to the worst performing employee in the wide span of control condition (thirteen participants, 9.03%) than in the low span condition (one participant, 0.68%).

Next, as mentioned, we only observed a negative effect of reporting frequency on the bonus of the best performing employee in the wide span of control condition. This result is consistent with the idea that some supervisors are reluctant to allocate exceptionally high bonuses and only do so when all the performance cues indicate that the employee should receive an exceptional bonus. Consistent with this explanation, we find that supervisors with a wide span of control are more likely to allocate the maximum bonus to the best performing employee in the low reporting frequency condition (44 participants, 60.27%) than in the high reporting frequency condition (30 participants, 42.25%). A Chi-squared test indicates this difference is significant (χ^2 =4.679, p = 0.031). For supervisors with a narrow span of control, the likelihood that they award the maximum bonus to the best performing employee is not significantly different (χ^2 = 0.1351, p = 0.713) between the low reporting frequency condition (fourteen participants, 19.18%) and the high reporting frequency condition (sixteen participants, 21.62%). These results suggest that when deciding whether to allocate an exceptionally high bonus supervisors do place disproportionate weight on a temporary dip in performance.

Finally, we examine if the data collected in the post-experimental questionnaire is consistent with the assumptions we made in our theory development. H1a and H1b are built on the idea that, as span of control widens, supervisors put more weight on the relative standing of the various employees under their supervision. Consistent with this assumption, participants indicate that their bonus decisions were more likely to be influenced by the relative annual net income (how the net income of each store manager compared to the net income of the other store managers) in the wide span of control condition (mean = 3.89, sd = 1.08) than in the narrow span of control condition (mean = 3.66, sd = 1.02, difference = - 0.24, t = 1.911, one-tailed p = 0.029).

3.4 Experiment Two

3.4.1 Design

We test the hypotheses related to employees' behavior (H4, H5, and H6) using an interactive lab experiment that, like Experiment One, has a 2 (span of control: narrow or wide) $\times$ 2 (reporting frequency: low or high) between-subjects design. Participants are randomly assigned to the role of employee or supervisor and are anonymously matched to form companies. Employees provide costly effort on a real-effort task for four periods. Employees' output is valuable to supervisors. After the last period, supervisors allocate a bonus to each employee in their company. In the narrow span of control condition, companies have two employees while in the wide span condition companies have five employees. In the low reporting frequency condition, supervisors receive aggregate performance reports after the final (fourth) period while in the high reporting frequency condition supervisors receive performance reports at the end of each period. We analyze two alternative measures of employee effort as dependent variables: the number of seconds employees worked on the real effort task, and employees' output.

Participants and Procedures

We conducted the experiment in CREED, the experimental economics laboratory of the University of Amsterdam. The participants were 165 members of the lab's subject pool who signed up after receiving an invitation to participate in our experiment. There were eight sessions, two for each condition, conducted over two days. Conditions were randomly assigned to sessions.[10] Each session was conducted with between 18 and 27 participants, depending on the number of show ups.

Participants' age varies from 17 to 58, with a mean of 22.01. In total, 79 participants (47.88%) are male and 86 (52.12%) are female. The majority (112 participants, 67.88%) indicated that their major was economics or business. Most participants (120, 72.72%) indicated they had at least some work experience, and 63 (38.18%) reported that they had a (part-time) job at the time of the experiment. On average, participants earned a total of €15.30 for about 50 minutes of their time.

Participants received initial verbal instructions (e.g. no communication, do not use phones) before moving to the computer lab where they were randomly assigned to a work station. We provided participants with a hardcopy set of instructions and gave them ten minutes to read these instructions before the computer task started. The computer task was programmed in z-Tree (Fischbacher, 2007). It started with a quiz containing understanding questions on which participants needed to score 100% before they could proceed to the main task. Once all participants passed the understanding quiz, they had the opportunity to familiarize themselves with the real effort task during a three-minute practice period. Next,

[10] Participants' characteristics do not differ between the two span of control conditions and between the two reporting frequency conditions with respect to age, gender, work experience, whether they study economics, self-reported risk preferences, self-reported mathematical skills, and whether they have evaluation experience. Results of a multiple linear regression do not reveal an association between the measured characteristics and assignment to one of the span of control conditions (F(8, 156) = 0.98, p = 0.451, R^2 = 0.05) or the reporting frequency conditions (F(8, 156) = 1.13, p = 0.340, R^2 = 0.06). These analyses suggest that random assignment was successful.

we randomly assigned participants to either the employee or supervisor role and matched participants to form companies. Participants then performed the main task, described in more detail below. The session ended with the participants completing a post-experimental questionnaire which contains items intended to provide insight into their' decision-making process during the main task, as well as measures of their risk preferences and demographics.

Out of the 165 participants, 43 were assigned the supervisor role and 122 were assigned the employee role (participants maintained their role and company throughout the session). Because our dependent variable is measured at the level of the employee, we tried to balance the number of employees between conditions. Because the number of employees per company varies depending on the span of control condition (two in the narrow span condition and five in the wide span condition), the number of companies in the two span of control conditions also varies. Specifically, in the narrow span of control condition, we collected data from 31 companies with a total of 62 employees, while in the wide span condition we collected data from twelve companies and 60 employees. Table 3.6 contains the specific number of (employee) observations in each experimental condition, which varies between 28 in the narrow span of control – low reporting frequency condition to 34 in the narrow span of control – high frequency condition.

Task and Manipulations

The main task consisted of four periods of 320 seconds (five minutes) during which the participants in the employee role performed a real effort task. The real effort task was the 'Word Encryption task with Double Randomization' ('WEDR task') developed by Benndorf et al. (2018) based on a similar task used by Erkal et al. (2011). In this task, participants encrypted three-letter words into numbers using a provided encryption table. The encryption table contained all letters of the alphabet, in random order, and a random number between 100 and 1,000 corresponding to each letter. For the three letters that form a word, employees needed to fill out the corresponding number from the encryption table.

After employees entered a correctly encrypted word, the computer generated a new word and a new encryption table. Employees could not proceed to the next word until they had correctly entered the current word.

We chose this task over similar real effort tasks used in previous studies because it minimizes the opportunities for learning behavior in repeated settings and because performance on this task is not substantially affected by participants' ability (Benndorf et al., 2018). We opted for a real effort task as opposed to a chosen effort task to enhance experimental realism (Charness et al., 2018). However, many real-effort tasks involve a loss of experimental control in the sense that researchers can no longer observe participants' cost of effort. To alleviate this issue we opted for a hybrid approach similar to the one used by Gächter et al. (2016). In each of the four work periods, employees lose one point out of an endowment of 300 points for every second spent working on the encryption task. Every period starts with a twenty-second grace period in which work is costless. After these twenty seconds, a STOP button appears on the employees' screen. Clicking this button immediately stops the encryption task. The seconds remaining at the moment the STOP button is clicked are transformed to points. Points, in turn, determine the payoff of the employee-participants from the experiment. Specifically, one point (second) is worth 0.5 Eurocents. Thus, for example, by clicking the STOP button immediately as it appears, a participant can secure $300 \times 0.5 = 150$ Eurocents (€1.50) in each work period. Employees who decide to stop working need to wait until the period is over and the experiment proceeds.

The role of the supervisor in the experiment was limited to assigning a discretionary reward to each employee in their company after the final (fourth) work period. For each employee, they could select a reward anywhere between zero and 3,000 points. To eliminate reneging temptations (Baiman & Rajan, 1995; Bol, 2008), the bonus selected by the supervisors was not deducted from the supervisors' payoff. Supervisors' pay depended on the output of their employees. In the narrow span condition, in which supervisors had two employees, each word correctly decoded by an employee contributed twenty points to their supervisor's total. In the wide span condition, in which supervisors had five employees, each word correctly decoded by an employee contributed eight points to their supervisor's total. Thus, the total expected number of supervisor points did not vary between conditions, however, an individual employee's contribution to this total was inversely proportional to the span of control.

Employees and supervisors also earned a fixed salary, which amounted to 300 points for the employees and 600 points for the supervisors. In summary, the payoff functions were as follows:

Employee: 300 + (1,200 – seconds worked) + Bonus, where the bonus was at the discretion of the supervisor and lay between zero and 3,000.

Supervisor: 600 + piece-rate × number of correctly encrypted words by the company's employees, where the piece rate was 20 in the narrow span condition and 8 in the wide span condition.

While the employees were working on the encryption task, supervisors could choose between sitting idle or also working on a similar encryption task. This choice, or their performance on the encryption task, did not affect the supervisors' payoff. Immediately after the supervisors assigned their discretionary rewards, employees were informed about their supervisor's reward decision. As emphasized in the instructions handout, employees were only informed about their own reward, and never learn the reward (or output) of the other employee(s) in the company.

Our main interest is in the level of – and variation in - employee effort. The experiment produced two measures of employee effort. The first was a "narrow" measure of effort in the agency theory sense: the number of seconds that subordinates spent working on the task. Baiman (1982) defines effort as a construct that is controllable by the employee, creates negative utility for the agent, and results in an increase in expected output. The number of seconds worked on the task satisfies these criteria. The second measure was employee performance, the number of correct answers provided in the aggregate of the four periods, which captured a broader notion of effort, such as the one provided by Bonner & Sprinkle (2002). This notion also takes into account other aspects besides effort duration, such as effort intensity, effort direction, and strategy development.

As indicated, span of control was manipulated by varying the number of subordinates that report to a specific supervisor. In the narrow (wide) span of control condition, each supervisor observed the performance - and assigned a reward to - two (five) employees. Reporting frequency was manipulated by varying how frequently supervisors were informed about the performance of their employees. In the low reporting frequency condition superiors were only informed about the aggregate performance of their subordinates after the final period. In the high reporting frequency condition superiors were informed about the performance of their subordinates after each of the four work periods. The reports they receive after each period listed the subordinates' performance in that period and the aggregate performance in the previous periods.

3.4.2 Results

Before testing our hypotheses we check whether our manipulations were effective. Results from manipulation checks included in the post-experimental questionnaire suggest that both the span of control manipulation and reporting frequency manipulation were salient to employees. First, we check the span of control manipulation by asking employee-participants to recall the number of employees (including themselves) who reported to their supervisor. Of the 62 participants in the narrow span condition 61 (98.4%) gave the correct answer and of the 60 participants in the wide span condition 57 (95.0%) gave the correct answer. To check the reporting frequency manipulation, we asked participants how often their performance was reported to their supervisor: 'continuously', 'after each period', 'only after the final period', or 'never'. 54 of the 58 participants in the low frequency condition (93.1%)

and 53 of the 64 participants in the high frequency condition (82.8%) answered this question correctly. We conclude that most participants were adequately able to recall the specifics of their condition. Since all participants also passed the understanding test in advance of the experiment, we test our hypotheses with the full sample. Notably, all results are qualitatively similar if we exclude the eighteen participants who failed at least one manipulation check.[11]

Table 3.6 contains descriptive statistics about employee effort in each condition, and overall. This table contains the mean and standard deviation of four variables: *TotalInput* is the aggregate number of costly seconds used (i.e., the total number of seconds used minus the 80 'free' seconds from the four grace periods), *SDInput* is the standard deviation of the number of costly seconds used in each of the four periods, *TotalOutput* is the aggregate number of correctly encrypted words in the four periods and *SDOutput* is the standard deviation of the number of correctly encrypted words in each of the four periods.

Hypothesis Tests

To test our hypotheses, we again rely on factorial ANOVAs and follow-up planned comparisons. As in Experiment One, we use two dummy variables labeled *WideSpan* and *HighFrequency* to identify conditions. First, H4 predicts that employees whose supervisor has a wider span of control will provide more effort. Panel A of Table 3.7 contains the ANOVA results for *TotalInput* and Panel B of Table 3.7 contains the ANOVA results for *TotalOutput*. In both panels, the main effect of *WideSpan* is insignificant (Panel A [*TotalInput*]: $F = 0.27$, $p = 0.603$; Panel B [*TotalOutput*]: $F = 0.00$, $p = 0.978$). Table 3.7 also shows that there is no significant effect of reporting frequency on *TotalOutput* and *TotalInput* (all $p > 0.100$) and that there is no significant interaction effect between *WideSpan* and *HighFrequency* on

[11]Seventeen participants failed one check and one participant failed both checks. All participants who failed the span of control manipulation check were off by one, i.e., they answered '3' instead of '2' or '4' or '6' instead of '5'. Regarding the reporting frequency manipulation check, two participants in the low frequency condition selected 'continuously', one participant selected 'after each period' and one participant selected 'never'. In the high frequency condition, one participant selected 'never' and ten participants selected 'after the last period only'.

Table 3.6: Descriptive Statistic: Effort in Experiment Two

Condition	Low Frequency			High Frequency			Overall		
	n = 28	mean	sd	n = 34	mean	sd	n = 62	mean	sd
Narrow Span of Control	TotalInput	886.04	378.38	TotalInput	972.91	244.73	TotalInput	933.68	312.53
	SDInput	44.73	55.26	SDInput	41.17	46.14	SDInput	42.78	50.07
	TotalOutput	74.07	32.35	TotalOutput	82.41	22.84	TotalOutput	78.65	27.62
	SDOutput	3.67	3.85	SDOutput	3.82	3.56	SDOutput	3.75	3.67
	n = 30	mean	sd	n = 30	mean	sd	n = 60	mean	sd
Wide Span of Control	TotalInput	900.37	327.28	TotalInput	897.77	334.20	TotalInput	899.07	327.95
	SDInput	49.05	54.54	SDInput	48.59	48.75	SDInput	48.82	51.29
	TotalOutput	80.33	29.32	TotalOutput	75.87	29.43	TotalOutput	78.10	29.22
	SDOutput	4.82	4.64	SDOutput	4.34	3.36	SDOutput	4.58	4.02
	n = 58	mean	sd	n = 64	mean	sd	n = 122	mean	sd
Overall	TotalInput	893.45	349.81	TotalInput	937.69	290.20	TotalInput	916.66	319.35
	SDInput	46.96	54.45	SDInput	44.65	47.15	SDInput	45.75	50.55
	TotalOutput	77.31	30.71	TotalOutput	79.34	26.13	TotalOutput	78.38	28.30
	SDOutput	4.26	4.28	SDOutput	4.06	3.45	SDOutput	4.16	3.85

This table presents the descriptive statistics of Experiment Two. It reports descriptive statistics on TotalInput (the aggregate number of costly seconds used), SDInput (the standard deviation of the number of costly seconds used across the four periods), TotalOutput (the aggregate number of correctly encrypted words in the four periods) and SDOutput (the standard deviation of the number of correctly encrypted words across the four periods) across the two span of control conditions and across the two reporting frequency conditions.

either dependent variable. Untabulated t-tests confirm that the effect of span of control on both effort measures is insignificant in both the low reporting frequency and high reporting frequency conditions (all $p > 0.100$). These results do not support H4.

Table 3.7: The Effects on the Effort Levels

Panel A - Analysis of Variance on Total Second Worked

Source	Partial SS	df	MS	F	p
Model	152,517.51	3	50,839.17	0.49	0.688
WideSpan	28,061.95	1	28,061.95	0.27	0.603
HighFrequency	53,891.09	1	53,891.09	0.52	0.472
WideSpan * HighFrequency	60,746.63	1	60,746.63	0.59	0.445
Error	12,187,494.00	118	103,283.85		
Total	12,340,012.00	121	101,983.57		

Panel B - Analysis of Variance on Total Performance

Source	Partial SS	df	MS	F	p
Model	1,376.43	3	458.81	0.57	0.638
WideSpan	0.61	1	0.61	0.00	0.978
HighFrequency	113.86	1	113.86	0.14	0.708
WideSpan * HighFrequency	1,244.52	1	1,244.52	1.54	0.218
Error	95,528.23	118	809.56		
Total	96,904.66	121	800.86		

Panel A of this table presents the results of a factorial ANOVA with the two manipulations as factors and TotalInput (the aggregate number of costly seconds used) as a dependent variable.
Panel B of this table presents the results of a factorial ANOVA with the two manipulations as factors and TotalOutput (the aggregate number of correctly encrypted words in the four periods) as a dependent variable.

To test H5 and H6 we run two more factorial ANOVAs with *SDInput* and *SDOutput* as dependent variables. The results are presented in Panel A and Panel B of Table 3.8 respectively. From Panel A of Table 3.8 it is clear that there is neither a main effect of *HighFrequency* ($F = 0.05$, $p = 0.829$) nor an interaction effect of *WideSpan* and *HighFrequency* ($F = 0.03$, $p = 0.867$) on *SDInput*. Similarly, Panel B of Table 3.8 shows that the main effect of *HighFrequency* ($F = 0.05$, $p = 0.815$) and the interaction effect of *WideSpan* and

HighFrequency (F = 0.20, p = 0.657) are not significant. These results are inconsistent with H5 and H6. In summary, none of our three hypotheses about the effects of span of control and reporting frequency on employee effort is supported by the data from Experiment Two.

Table 3.8: The Effects on the Standard Deviation in Effort

Panel B - Analysis of Variance on the Standard Deviation In Seconds Worked

Source	Partial SS	df	MS	F	p
Model	1,310.69	3	436.90	0.17	0.918
WideSpan	1,045.50	1	1,045.50	0.40	0.528
HighFrequency	122.30	1	122.30	0.05	0.829
WideSpan * HighFrequency	73.23	1	73.23	0.03	0.867
Error	307,889.91	118	2,609.24		
Total	309,200.60	121	2,555.38		

Panel B - Analysis of Variance on the Standard Deviation In Performance

Source	Partial SS	df	MS	F	p
Model	24.75	3	8.25	0.55	0.649
WideSpan	21.25	1	21.25	1.42	0.237
HighFrequency	0.82	1	0.82	0.05	0.815
WideSpan * HighFrequency	2.98	1	2.98	0.20	0.657
Error	1,771.16	118	15.01		
Total	1,795.91	121	14.84		

Panel A of this table presents the results of a factorial ANOVA with the two manipulations as factors and SDInput (the standard deviation of the number of costly seconds used across the four periods) as a dependent variable.
Panel B of this table presents the results of a factorial ANOVA with the two manipulations as factors and SDOutput (the standard deviation of the number of correctly encrypted words across the four periods) as a dependent variable.

3.5 Discussion and Conclusion

We examine how discretionary award allocations by supervisors are affected by two features of the organizational control system: the reporting structure (a supervisor's span of control) and the reporting frequency. Using an online case-based experiment, we find that strong performers receive higher rewards and weak performers receive lower rewards from supervisors with wider spans of control. Contrary to our expectations, we did not find that reporting frequency affects reward allocations. In addition, we examine whether employees anticipate supervisors' reward allocations, and adjust their effort levels based on their supervisor's span of control and the frequency with which their performance is reported. Using an interactive laboratory experiment, we find no evidence that employee effort is affected by span of control and reporting frequency.

Despite the lack of support for many of our hypotheses, we believe our study makes several contributions to the accounting literature. First, we contribute to the discretionary evaluation literature by investigating how span of control influences performance evaluations. Previous literature finds mixed results when examining the effect of span of control on evaluations (Ellington & Wilson, 2017; Judge & Ferris, 1993; Lahuis & Avis, 2007; O'Neill et al., 2012). Our study helps explain these mixed results by providing evidence that span of control affects evaluations differently depending on an employee's standing among their peers. Moreover, these findings should interest companies considering to "flatten their organization" (Hannan et al., 2010) because they can help controllers anticipate how discretionary evaluations will change when span of control widens.

Second, we further contribute to the discretionary evaluation literature by examining whether employees adjust their effort levels in anticipation of how different control systems will affect supervisors' evaluations. In the case-based experiment, we found that supervisors change their evaluation patterns when span of control widens. If employees accurately anticipate how supervisors will evaluate them, as it is sometimes assumed in the literature (e.g. Baiman & Rajan, 1995), we should have observed that employees increase their effort as span of control widens in our lab study. However, our results show no evidence of such an effect.

Researchers have found that employees anticipate how a different control system will affect their evaluation in some situations (Chan, 2018) while they do not in other situations

(Arnold et al., 2018; Ghita, 2021b). Based on what we currently know, it is difficult to propose a theory that would explain why changing the control system did not alter employee behavior in our setting while it did in other settings (Chan, 2018). However, we believe the evidence we provide represents an important step towards developing a more comprehensive theory regarding how employees behave in a discretionary evaluation setting.

Third, we contribute to the emerging literature on the effects of changing reporting frequency within the firm. Our findings complement the conclusions of a recent study on the subject (Hecht et al., 2020). While Hecht et al. (2020) find that a higher reporting frequency decreases employee motivation in a setting where performance depends primarily on employee ability, we find no effect of reporting frequency on employee behavior in a setting where performance depends primarily on employee effort. Our results expand our understanding of how reporting frequency affects employee motivation and suggest that this effect may depend on the type of task, specifically on the relative importance of effort and ability as antecedents of employee performance.

Our study is subject to several limitations that provide opportunities for future research. First, because we used the experimental method, we were not able to capture all consequences of changing supervisors' span of control in our design. Future research could further investigate how other consequences of widening supervisors' span of control affect discretionary evaluations. For example, a wider span of control likely makes it more costly for supervisors to observe each employee's performance and circumstances (Gong et al., 2019; O'Neill et al., 2012). This additional cost could cause supervisors to gather insufficient information about their employees and to become more lenient in their evaluations (Bol, 2011; Gong et al., 2019; Maas & Verdoorn, 2017). However, because supervisors are concerned with relative fairness, a wider span of control could also make supervisors more concerned about providing accurate evaluations which, in turn, could increase the amount of information supervisors gather about their employees (Maas et al., 2012).

Second, while we believe the difference in task type (effort or ability dominated) explains why our conclusions differ from the study of Hecht et al. (2020), we did not explicitly manipulate task type. Therefore, other differences in our designs could be causing the differences in results. Future research could, for example, manipulate employee perception about the relative importance of effort and ability as antecedents of performance to expand our understanding of how reporting frequency affects employee behavior.

Chapter 4

Do Supervisors Reward Observable Luck?

4.1 Introduction

One of the main difficulties in designing efficient incentive systems is filtering out the effect of luck from employees' compensation (Baker et al., 1994; Bol, 2011; Gibbs et al., 2004). A common and theoretically viable solution to this issue is allowing supervisors discretion in evaluating their direct subordinates (Baker et al., 1994; Bol, 2008). Discretion allows supervisors to incorporate relevant but non-contractible information when evaluating employees. This informational advantage of discretionary evaluations relative to explicit contracts permits supervisors to reduce the influence of luck on employee compensation (Baker et al., 1994; Bol, 2008; Gibbs et al., 2004). For example, a supervisor may learn that an unforeseeable natural catastrophe hurt a manager's performance and appropriately adjust compensation so that the manager is not punished for their bad luck. Supervisors are expected to use the information available to them to filter out luck from evaluations because it is profitable for them to do so (Baiman & Rajan, 1995; Baker et al., 1994) and because they find it fair to reward employees for their contribution (Bol et al., 2016; Chan, 2018;

Maas et al., 2012).

Despite these strong theoretical predictions, supervisors sometimes seem to reward good luck and punish bad luck in their discretionary evaluations (Bol et al., 2015; Merchant, 1987). Previous literature argues that supervisors reward luck because the effect of luck on performance is generally unobservable (Merchant & Otley, 2006). I examine whether supervisors reward luck even when it is *observable*, that is, the effects of luck are known and can be quantified. I expect supervisors will reward observable good luck and punish observable bad luck. This might be the case because supervisors are influenced by fairness concerns when evaluating employees (Bol et al., 2016; Chan, 2018; Maas et al., 2012) and because some supervisors might find it fair to reward luck (Cappelen et al., 2007; Cushman et al., 2009). If supervisors reward observable luck, allowing supervisor discretion may not always lower the impact of luck in employee compensation even if supervisors have access to additional non-contractible information. Thus, one of the main theoretical benefits of implementing discretionary evaluations may be weaker than conventional economic reasoning suggests (Baker et al., 1994).

I also examine how employees' behavior differs when supervisors reward observable luck. Accounting researchers consistently find that supervisors' opportunism, cognitive limitations, and social preferences prevent supervisors from evaluating employees in a manner that analytical research would consider optimal (Arnold & Tafkov, 2019; Bailey et al., 2011; Bol et al., 2016; Bol & Smith, 2011; Krishnan et al., 2005; Lipe & Salterio, 2000). Although supervisors' divergences from theoretically optimal evaluations are well documented, we know relatively little about how these distorted evaluations affect employee behavior (Ahn et al., 2010; Arnold et al., 2018; Berger et al., 2013; Bol, 2011). It is important to examine the effect of these deviations on employee behavior because in some situations the theoretically imperfect evaluations appear to motivate employees (Bol, 2011) while in other situations the theoretically imperfect evaluations appear to demotivate employees (Ahn et al., 2010;

Arnold et al., 2018).

The controllability principle asserts that supervisors should not reward observable luck (Merchant, 1987). Rewarding luck weakens the link between employee contribution to company value[1] and their compensation, which can decrease employee contribution (Bol, 2008; Bol & Smith, 2011; Cadsby et al., 2019; Vroom, 1964). Despite this, supervisors could reward observable luck for at least two reasons. First, supervisors want to provide fair evaluations (Maas et al., 2012) and could consider it fair to reward luck. For example, if a supervisor gets a higher bonus because an employee had good luck, the supervisor could consider it fair to split some of the surplus of that luck with the employee. This logic is compatible with a consequentialist view of fairness: an act is right or wrong solely on the basis of whether it maximizes good outcomes (Uhlmann et al., 2015). Diverse streams of literature find that many people judge the fairness of actions in a manner consistent with this consequentialist view of fairness (Cushman et al., 2009; Hannan, 2005; Rubin & Sheremeta, 2015). Fair-minded supervisors who have consequentialist views find it fair to reward employee performance, which is determined by both employee contribution and by luck, instead of employee contribution. As a result, these supervisors could partially reward observable luck because they find it fair to do so.

Second, supervisors could reward observable luck because they think their employees have consequentialist views of fairness. Supervisors will try to provide evaluations that the employees perceive as fair because such evaluations motivate employees (Chan, 2018; Fehr et al., 2009) and decrease the chance of costly confrontation with employees (Bol, 2011; Bol et al., 2016). Moreover, employees' self-serving fairness concerns could cause supervisors to integrate observable luck asymmetrically, that is, supervisors could reward good luck more than they punish bad luck. This is because employees have self-serving fairness concerns and

[1]Employees can contribute to company value through their effort, ability, and knowledge (Libby & Luft, 1993).

likely consider it fair to be rewarded for good luck and unfair to be punished for bad luck (Arnold et al., 2018; Asay et al., 2019; Bol & Smith, 2011; Feather, 1999; Gibbs et al., 2004; Hannan, 2005).

The idea that supervisors reward observable luck is consistent with a large body of literature in accounting and psychology on outcome bias (Baron & Hershey, 1988; Mertins et al., 2013). In addition, many different streams of literature consistently find that people reward luck when evaluating others (Akbaş et al., 2019; Brownback & Kuhn, 2019; Cushman et al., 2009; Rubin & Sheremeta, 2015). Moreover, previous accounting literature has also examined if supervisors asymmetrically integrate luck in evaluations (Bol & Smith, 2011; Hannan, 2005). However, these findings arise in a context that does not include an economic incentive for evaluators to ignore luck. This is because previous studies either used a hypothetical decision to test their hypotheses (Bol & Smith, 2011; Mertins et al., 2013) or studied a setting where a supervisor and an employee interacted only once (Brownback & Kuhn, 2019; Hannan et al., 2012; Rubin & Sheremeta, 2015). In contrast to the evaluators from previous studies, middle-level supervisors have an incentive to ignore luck because doing so could result in higher employee contribution (Bol, 2008; Cadsby et al., 2019; Vroom, 1964). Examining how middle-level supervisors reward observable luck is therefore informative given that decision-makers are more likely to ignore observable luck when they have an incentive to do so (Rand et al., 2015).

It is unclear ex-ante how employees will react if supervisors incorporate observable luck in their evaluations. Rewarding luck weakens the link between employee contribution and their compensation which, in turn, can decrease employee contribution (Bol, 2008; Cadsby et al., 2019; Vroom, 1964). However, employees likely do not choose their contribution solely to maximize their compensation (Charness, 2004; Falk & Fischbacher, 2006). Instead, employees also increase their contribution after evaluations that they consider fair (Bol, 2011; Colquitt et al., 2001). Employees with consequentialist views of fairness could increase their contribution when supervisors reward observable luck because they consider it fairer to be rewarded for luck. Moreover, employees could expect supervisors to reward them for good luck and shelter them from bad luck (Bol & Smith, 2011; Hannan, 2005). Employees could decrease their contribution as a form of punishment against supervisors who do not reward good luck and, therefore, do not fulfill this expectation.

I conduct an interactive, multiperiod experiment with 258 participants from Amazon Mechanical Turk in which randomly matched pairs of participants assume the role of either a supervisor or an employee. Employee contribution and luck determine profit. The employee can increase expected company profit by increasing their contribution. Company profit is valuable to the supervisor. The supervisor assigns a bonus to the employee based on information produced by the reporting system. The supervisor does not pay the bonus out of his or her own pocket. After each evaluation, the employee can confront the supervisor. A confrontation reduces the payoff of both the employee and the supervisor. The reporting system can produce two performance measures: *employee contribution level*, a perfect measure that captures contribution and is unaffected by luck and *the profit level*, a noisy measure that partially captures contribution and is influenced by luck. I manipulate the information presented to the supervisors: supervisors either observe both the profit level and the employee contribution level or they only observe the employee contribution level.

I find supervisors reward observable luck. Results from the post-experimental questionnaire confirm that supervisors' tendency to reward observable luck reflects fairness concerns. A majority of participants consider it fair that luck is incorporated in discretionary evaluations. Contrary to my initial prediction, I find supervisors punish bad luck more than they reward good luck. Although supervisors anticipate employees find it fairer to be rewarded for good luck than to be punished for bad luck, supervisors do not integrate employee self-serving fairness concerns when employee contribution is low. This might be the case because supervisors integrate employees' fairness concerns in evaluations as a form of reciprocity. If this is true, supervisors would not feel the need to reciprocate, that is, offer evaluations that their employees would consider fair when employee contribution is low. In line with this, I find supervisors care less about offering evaluations that the employees would consider fair when employee contributions are lower.

Employees' contribution is lower when supervisors reward observable luck but only after employees learn how supervisors evaluate them through repeated interactions. I perform an additional analysis that examines if employees with consequentialist fairness views react negatively to not being rewarded for luck. I find that, when supervisors do not reward luck, employees with consequentialist views of fairness confront supervisors more when they have good luck than when they have bad luck.

I contribute to the literature on discretionary evaluations in four ways. First, previous studies find that supervisors' opportunism, cognitive limitations, and social preferences prevent supervisors from providing theoretically optimal evaluations (Bailey et al., 2011; Bol et al., 2016; Bol & Smith, 2011; Chan, 2018; Krishnan et al., 2005; Lipe & Salterio, 2000). These distortions diminish the intended contracting benefits of discretionary evaluations (Baiman & Rajan, 1995; Baker et al., 1994; Feltham & Xie, 1994; Rajan & Reichelstein, 2006). I contribute to this line of research by showing that supervisors reward observable luck. As a consequence, allowing supervisor discretion may not always lower the impact of luck on employee compensation even if supervisors have access to additional non-contractible information.

Second, I expand our understanding of how fairness concerns influence discretionary evaluations. While it is well established that fairness concerns affect discretionary evaluations (Arnold & Tafkov, 2019; Bol, 2011; Bol et al., 2015; Chan, 2018; Maas et al., 2012), we do not have a comprehensive theoretical framework that explains which evaluations are considered fair by supervisors and employees (Cappelen et al., 2007; Konow, 2003). Accounting researchers have argued that a sense of fairness causes supervisors to only hold employees accountable for factors that employees can immediately control (Arnold & Tafkov, 2019; Bol, 2011; Chan, 2018; Maas et al., 2012). In this paper, I expand our understanding of how controllability influences fairness perception by showing that some supervisors and employees consider it fair that completely uncontrollable factors are incorporated into evaluations.

Third, I expand our understanding of how employees' self-serving fairness perceptions influence discretionary evaluations (Arnold et al., 2018; Arnold & Tafkov, 2019; Bol et al., 2016). Employees' self-serving fairness perceptions can lead to leniency bias in discretionary evaluations partially because supervisors want to avoid confrontations with their employees (Arnold & Tafkov, 2019; Bol, 2011; Bol et al., 2016; Deason et al., 2018; Moers, 2005). I find that although supervisors anticipate employees' self-serving fairness perceptions and confrontations with the employees are costly, supervisors do not indiscriminately integrate employees' fairness perceptions in discretionary evaluations. Specifically, supervisors do not integrate these self-serving fairness perceptions when employees have a low contribution to company value. This suggests supervisors incorporate employee self-serving fairness perceptions as a form of reciprocity towards employees who already have a high enough contribution to trigger reciprocity.

Fourth, I examine how employees change their behavior when supervisors reward observable luck. Rewarding luck results in a lower employee contribution after employees learn how supervisors evaluate them. When supervisors do not reward luck, I find employees with consequentialist views of fairness confront supervisors more when they have good luck than when they have bad luck. This indicates some employees confront supervisors for not rewarding luck. Supervisors who initially base evaluations exclusively on employee contribution could, through repeated interactions with these employees, integrate luck into their evaluations to avoid costly confrontations. Practitioners or researchers who design interventions that aim to increase the weight of employee contributions in discretionary evaluations (Berger et al., 2013; Bol et al., 2018, 2016; Demeré et al., 2019) should therefore also consider whether employees find it fair to be rewarded for luck and what control mechanisms could change employee fairness perceptions towards a less consequentialist view of fairness.

4.2 Theory

I first review related literature on the factors affecting middle-managers' discretionary evaluations that are relevant to the current research question. Then, I develop hypotheses related to whether supervisors will incorporate observable luck in their evaluations. Finally, I develop a research question regarding how employees will change their contribution to company value when supervisors reward observable luck.

4.2.1 Background

In many organizations, middle-level managers have discretion over the evaluation of their direct subordinates. Organizations provide supervisors with this discretion because it allows supervisors to incorporate non-contractible information in employee compensation (Bol, 2008) and because explicit contracts are often prohibitively costly to write (Choi et al., 2016).

A considerable body of management accounting research examines how supervisors use their discretion and finds that at least three considerations influence discretionary evaluations. First, supervisors use their evaluation decisions to motivate higher employee contribution to company value, because higher employee contribution likely benefits supervisors (Baiman & Rajan, 1995; Bol, 2008). For example, higher employee contribution might increase departmental performance, which in turn might increase supervisors' compensation and promotion opportunities (Bol et al., 2016). Second, supervisors' social preferences affect their evaluations. Notably, supervisors give evaluations that they find fair (Chan, 2018; Fehr & Schmidt, 2001; Maas et al., 2012). Third, supervisors use evaluations to minimize personal costs, such as confrontation costs (Prendergast & Topel, 1993). Confrontations with employees are costly because they may, for example, require supervisors to spend additional time collecting information that justifies their assessment. Therefore, to minimize potential confrontation costs, supervisors prefer providing ratings that satisfy employees (Bol et al., 2016). Employees are more satisfied with their evaluation when they perceive their evaluation as fair (Colquitt et al., 2001). Therefore, when making evaluation decisions, supervisors consider the fairness of the evaluation from the perspective of employees. Employee fairness perceptions are partly driven by receiving adequate compensation for their contribution and partly driven by self-serving considerations (Arnold et al., 2018; Feather, 1999; Miller & Ross, 1975). These self-serving considerations are the tendency of individuals "to conflate what is fair with what benefits oneself" (Babcock et al., 1996), resulting in individuals developing fairness rules biased towards increasing their own payoffs (Arnold et al., 2018; Deason et al., 2018).

4.2.2 Hypotheses

Supervisors Reward Observable Luck

The controllability principle states that employees should only be held accountable for what they can at least influence (Merchant & Otley, 2006). This principle suggests that supervisors should not reward observable luck because doing so decreases employee motivation. Supervisors who reward luck weaken the link between employee contribution and their compensation. This can decrease employee motivation in two ways. In short, rewarding luck can lower employee motivation because it exposes employees to more risk when they increase their contribution and because employees could find it unfair to be rewarded for factors they cannot control.

First, rewarding luck can expose employees to more risk when employees increase their contribution to company value (Baker & Jorgensen, 2003; Cadsby et al., 2016, 2019; Zubanov, 2012) and, therefore, decreases the utility the employees gain by increasing their contribution.[2]

The next example illustrates this point. Consider the case of a supervisor who evaluates salespeople to determine their bonuses. The salespeople are effort-averse. Most low-effort days result in low sales. This means that both the probability of high sales and the variation in sales are low on a low-effort day. When salespeople exert effort, they make some sales and miss others because sales also depend on luck such as clients' financial situation, the number of competing offers, etc. Therefore, both the probability of high sales and the variation in sales are higher when salespeople exert effort than when they decide to shirk (Baker & Jorgensen, 2003; Cadsby et al., 2019).[3]

The supervisor initially considers only rewarding salespeople based on their sales and not filtering out observable luck. This is a good motivational mechanism for some salespeople because, by exerting effort, salespeople can increase the probability of new sales. Although attempting new sales increases the expected bonus, it also imposes additional risk on the salespeople. Shirking, in contrast, leads to a lower expected bonus but imposes less risk on the salespeople. To minimize their risk, risk-averse salespeople will sacrifice some of the increases in the expected bonus gained from attempting some sales. Therefore, if the supervisor reduces the weight of luck in evaluations by, for example, integrating the number of clients contacted in the bonus decision, risk-averse salespeople will attempt more sales.

To motivate employees, supervisors should filter out observable luck, because doing so maximizes employee utility when employees have high contributions. When supervisors have discretion in evaluations, employees are uncertain about how supervisors will evalu-

[2]This assumes that most people are risk-averse (Holt & Laury, 2002) and that companies cannot perfectly select employees based on risk-preferences through their hiring process.

[3]Differently from studies that consider effort and luck are independent of one another (luck is an additive term), I consider a setting where effort and luck interact. Effort likely affects the influence of luck on performance in many real-world settings (Baker & Jorgensen, 2003; Cadsby et al., 2019). Additionally, if I treated luck as an additive term, rewarding luck should not, absent fairness considerations, decrease employee contribution (Sloof & Van Praag, 2010). By studying a setting where effort and luck interact, I provide supervisors with an additional incentive to ignore luck.

ate them (Bol, 2008; Choi et al., 2016; Gibbons & Henderson, 2012). This uncertainty reduces employee incentives to increase their contribution (Bol, 2008). By offering bonuses that maximize employee utility when employee contribution is high, supervisors increase the probability that bonuses are high enough to compensate employees for the cost of their contribution and for the uncertainty accepted by working.

Second, employees can decrease their contribution when supervisors reward observable luck because they find evaluations that are based on factors they cannot control unfair (Kelly et al., 2015; Maas et al., 2012). Employee fairness perceptions about their evaluations have an important impact on employee motivation (Bol, 2011; Colquitt et al., 2001). Some people's notion of fairness is consistent with the just deserts theory of fairness (Arnold & Tafkov, 2019; Chan, 2018; Falk et al., 2008; Falk & Fischbacher, 2006; Maas et al., 2012). Employees with just deserts fairness views believe that their compensation should only be based on factors they can control (e.g. how hard they worked). If supervisors reward observable luck, these employees will find their evaluations unfair and decrease their contribution.

Despite these two benefits of filtering out luck from evaluations, supervisors could still reward observable luck because they find it fair to do so. Some people have a consequentialist view of fairness: they believe an act is right or wrong solely on the basis of whether it maximizes good outcomes (Uhlmann et al., 2015). Different streams of literature find results consistent with the idea that some people have a consequentialist view of fairness. For example, Cushman et al. (2009) find that people are more likely to positively reciprocate a bad intention that resulted in a good outcome than a good intention that resulted in a bad outcome. Experiments in risk-taking on others' behalf, redistribution, and charitable giving find that people reward others based on both their choices and their luck (Cappelen et al., 2013; de Oliveira et al., 2017; Falk et al., 2008; Gurdal et al., 2013; Pan & Xiao, 2016).

Supervisors with consequentialist views of fairness find it fair to reward performance instead of employee contribution to that performance. Because performance is influenced by luck, these supervisors likely find it fair to partially reward luck.[4] Moreover, supervisors could anticipate that some employees have consequentialist fairness views and would perceive evaluations based exclusively on their contribution as unfair. Employees who consider they have received an unfair evaluation might decrease supervisors' payoff by lowering their contribution (Bol, 2011; Fehr et al., 2009) or by confronting the supervisors (Bol et al., 2016). Supervisors can motivate higher employee contribution and fewer confrontations by offering evaluations that are perceived as fair by employees. Therefore, if supervisors find it fair to reward luck or if supervisors anticipate that some of the employees find it fair to be rewarded for luck, they will incorporate observable luck into discretionary evaluations.

H1: Supervisors incorporate observable luck in discretionary evaluations

The idea that supervisors reward luck is consistent with a large body of literature in accounting and psychology on outcome bias (Baron & Hershey, 1988; Mertins et al., 2013).[5] This literature is mainly concerned with examining how supervisors rate the quality of a decision depending on the outcome of that decision. For example, Brazel et al. (2016) find that auditors are evaluated based on the outcome (whether a material misstatement was found) of their decision to perform additional tests and not the validity of the decision itself. Also relevant to the current study is a separate recent stream of literature that uses an economics-based approach to examine whether principals reciprocate based on outcomes or agent contribution in an incomplete-contract setting (Brownback & Kuhn, 2019; Hannan, 2005; Rubin & Sheremeta, 2015).

Results from the previous studies do not automatically translate to a setting where

[4] Accounting researchers have argued that a sense of fairness causes some middle-level supervisors to only hold employees accountable for factors that employees can immediately control (Arnold & Tafkov, 2019; Bol, 2011; Chan, 2018; Maas et al., 2012). This conclusion can seem to suggest that all middle-level supervisors have beliefs consistent with a just deserts theory of fairness. However, because different people have different views about what is fair (Cappelen et al., 2007; Konow, 2003), discretionary evaluations can be influenced by both the just deserts view and by the consequentialist view.

[5] The executive compensation literature has also examined whether executives are rewarded for luck (Bertrand & Mullainathan, 2001; Garvey & Milbourn, 2006). However, the evaluation context of an executive is different than the evaluation context of a non-executive employee.

middle-level supervisors evaluate employees because evaluators from previously cited studies were not motivated to reward employee contribution and ignore luck. The outcome bias literature from psychology and accounting uses an unincentivized hypothetical decision to test if outcome bias influences evaluations (Brazel et al., 2016; Mertins et al., 2013; Sezer et al., 2016).[6] The economics studies that examined how luck influences evaluations (Brownback & Kuhn, 2019; Rubin & Sheremeta, 2015) model the evaluator as the owner of the firm who has to pay the employee bonus out of their own pocket and only interacts with the employee once. In contrast, middle-level managers do not pay employee bonuses themselves (Bol et al., 2016; Prendergast & Topel, 1993) and usually evaluate employees for multiple periods. Because owner-evaluators are motivated to minimize the bonus paid to employees, owner-evaluators have an incentive to renege and give employees the minimum bonus regardless of employees' contribution level. Moreover, because owner-evaluators only interact with the employee once, they are not motivated to choose a bonus that will motivate the employee in future periods. In contrast, middle managers face no reneging temptations (Bol, 2008) and can influence employees' choices in future periods through their bonus decisions. Therefore, middle managers have stronger incentives than owner-evaluators to choose an employee bonus that is most likely to maximize employee contribution. If, as suggested by the controllability principle, supervisors believe that exclusively rewarding employee contribution maximizes their payoff, supervisors will be motivated to ignore their consequentialist fairness views and exclusively reward contribution.

[6]Most of the outcome bias studies use the following definition for outcome bias: people perceive the same decision to be lower in quality when it leads to a bad outcome rather than a good outcome, all else being equal (Sezer et al., 2016). Therefore, the supervisors from the outcome bias studies are asked to judge the quality of a decision. However, when evaluating how much an employee has contributed to company value, supervisors are not judging the quality of a decision because the best employee decision is to maximize contribution. For example, a supervisor always knows that employee shirking decreases company value whereas a supervisor does not know whether an additional auditing test should have been performed. Therefore, the current study also differs from previous outcome bias studies because the supervisors cannot learn any useful information about the decision quality when observing the outcome of that decision (Mertins et al., 2013; Weber et al., 2001).

Supervisors' beliefs about what maximizes their payoff are important because when people believe that ignoring luck will result in higher payoff for themselves, they are less likely to reward luck. Rand et al. (2015) examined cooperation in a repeated prisoner's dilemma game where participants choose an intention (cooperate or defect). The other player could perfectly observe the intention but the final action also depended on luck. Differently than all other previous cited studies, participants could potentially gain more by reacting to intentions and ignoring luck. They find that a vast majority of participants conditioned their choice on intentions and ignored luck.

Because supervisors could believe that ignoring luck when evaluating employees maximizes their payoff, one could expect that, similar to Rand et al. (2015), supervisors will base their evaluations exclusively on employee contribution. However, it is unlikely that all supervisors will believe that ignoring observable luck will maximize their payoff. In some settings, supervisors believe that incorporating luck in their evaluations maximizes employee contribution (Bol et al., 2016; Giraud et al., 2008; Merchant, 1987). Employees could have consequentialist fairness views and perceive evaluations based exclusively on contribution as stringent and unfair (Hannan, 2005). Given there is uncertainty about how employees react to evaluations based exclusively on contribution, supervisors will likely be unsure whether this evaluation strategy maximizes their payoff. Therefore, supervisors will integrate observable luck in their evaluation partly because they consider it fair to do so and partly because their employees might consider it fair to do so.

Supervisors Reward Observable Good Luck More than They Punish Observable Bad Luck

Employees judge fairness differently than their supervisors partially because people are self-serving when they form fairness judgments (Arnold et al., 2018; Feather, 1999). Employees could consider that evaluations determined exclusively by contribution are unfair, especially in situations in which this evaluation strategy results in lower evaluations for them.

Employees with self-serving fairness perceptions likely consider it fair to be rewarded for good luck and unfair to be punished for bad luck (Asay et al., 2019; Bol & Smith, 2011; Gibbs et al., 2004; Hannan, 2005). Employees punish supervisors for perceived unfair evaluations by either confronting them or by reducing their contribution in future periods (Bol, 2011; Bol et al., 2016). Therefore, to avoid this, supervisors could asymmetrically reward luck such that they reward good luck more than they punish bad luck.

H2: Supervisors reward observable good luck more than they punish observable bad luck

It is not obvious that supervisors will asymmetrically reward luck. Although it is relatively well established that employees have self-serving fairness perceptions (Asay et al., 2019; Bol & Smith, 2011; Gibbs et al., 2004; Hannan, 2005), it is less clear if these fairness perceptions will be integrated into discretionary evaluations. In order for supervisors to reward good luck more than they punish bad luck, supervisors need to anticipate employees' self-serving fairness perceptions and they need to believe that it is beneficial for them to integrate employees' fairness perceptions in evaluations. The repeated interaction between supervisors and employees could, for example, cause supervisors to believe that they are better off ignoring luck completely and only rewarding employee contribution (Rand et al., 2015).

Employee Contribution when Supervisors Reward Observable Luck

It is unclear ex-ante how rewarding observable luck will affect employee contribution. On the one hand, rewarding luck decreases the link between employee contribution and their compensation which, in turn, can decrease employee contribution (Bol, 2008; Cadsby et al., 2019; Vroom, 1964). In line with the view that rewarding luck can decrease employee contribution, Rubin & Sheremeta (2015) find that, in a single-interaction setting, agents are less likely to provide the desired amount of effort when principals reward luck.[7]

On the other hand, rewarding luck could improve employees' fairness perceptions which, in turn, can increase employee contribution (Bol, 2011; Charness, 2004; Falk & Fischbacher, 2006). Employees with consequentialist views of fairness likely consider evaluations that incorporate observable luck as fairer. Moreover, if supervisors integrate good luck more than they punish bad luck, employees can perceive the resulting higher evaluations as a gift from their supervisors. As a result, employees who care about fairness could reciprocate and increase their contribution when evaluations are partially based on observable luck (Bol, 2011; Charness, 2004; Falk & Fischbacher, 2006). In line with the view that rewarding luck can increase employee contribution, Hannan (2005) finds that, in a single-interaction setting, agents increase their effort when principals share the benefits of good luck with them.

Because I am not able to predict whether the possible increase in positive reciprocity dominates the weaker incentives caused by incorporating observable luck, I present a research question.

RQ: Is employee contribution lower when supervisors incorporate observable luck in discretionary evaluations?

4.3 Method

In my experiment, participants are randomly assigned to the role of either employee or supervisor. One supervisor is anonymously matched with one employee, and each pair remains matched for eight periods. In each period, the employee first chooses how much to contribute (employees perform a chosen effort task). Employee contribution and luck determine company profit. Profit is valuable to the supervisor. After each employee-contribution decision, the supervisor decides how much bonus the employee receives. The employee then

[7]Given that Rubin & Sheremeta (2015) study a single-interaction setting and employee reciprocity plays a limited role in determining effort in single-interaction settings (Fisher et al., 2015), it is unclear whether rewarding luck will decrease employee contribution when supervisors and employees interact for multiple periods.

observes the bonus and decides whether to confront the supervisor. A confrontation reduces the payoff of both the employee and the supervisor.

Between subjects, I manipulate the information presented to the supervisor when making the bonus decision at two levels. The reporting system either reports (1) company profit and employee contribution (the Outcome+Contribution condition) or (2) only employee contribution (the Contribution-Only condition). Within subjects, company profit (high or low) is quasi-randomly determined every period.[8]

To test how observable luck influences evaluations (H1 and H2), I examine how the profit level influences bonuses in the Outcome+Contribution condition. In the Outcome+Contribution condition, supervisors can directly observe employee contribution so the profit level only contains information about luck. If bonuses are influenced by profit after controlling for employee contribution, I can conclude that supervisors reward observable luck. To examine whether employees' contribution differs when supervisors reward observable luck (RQ), I compare employee contribution between the Outcome+Contribution and Contribution-Only conditions.[9] Supervisors in the ContributionOnly condition cannot reward luck because they do not observe luck. Therefore, the ContributionOnly condition allows me to observe employee contribution when supervisors do not reward observable luck and have access to the same amount of information about employee contribution as supervisors in the Outcome+Contribution condition.

4.3.1 The Task

Every period, the computer simulates drawing a random ball from a bag that contains winning and losing balls. Company profit is high (low) if a winning (losing) ball is drawn. In the first decision of each period, employees decide how much effort they want to invest,

[8] Employee contribution influences how much influence luck has in determining company profit. However, I can attribute differences in bonus to differences in luck in regressions that control for employee contribution (Brownback & Kuhn, 2019; Stock & Watson, 2015). Therefore, I can reliably assess causality.

[9] I can only examine how employees react when supervisors reward observable luck if supervisors reward observable luck in the Outcome+Contribution condition.

i.e., they perform a chosen effort task. Employees can spend a part of their endowment to buy winning balls. Buying winning balls replaces losing balls with winning balls. The bag initially contains two winning balls and eight losing balls. Employees can choose to do nothing (buy no winning balls) or buy up to five winning balls.[10] For example, if an employee chooses to buy three winning balls, there will be five winning balls and five losing balls in the bag (the chance of a high profit is 50%). Each winning ball reduces the payoff of the employee by twenty points. Table 4.1 describes how this employee decision affects the probability that company profit is high and the employee's endowment. After deciding how many winning balls to buy, employees are informed about the profit level for that period.

Table 4.1: Consequences of Employee Contribution Decision

Number of winning balls bought	0	1	2	3	4	5
Probability that profit is high	20%	30%	40%	50%	60%	70%
Cost to the employee	0	20	40	60	80	100

After the employee contribution decision, the supervisor assigns a bonus to the employee, between zero and 200 points. The bonus comes from a fixed bonus pool and is not paid out of the supervisor's pocket. I use a fixed bonus pool because a variable bonus pool would implicitly reward employees for luck and therefore add a confounding effect to the analysis.[11] The supervisor's payoff depends on company profit. Specifically, the supervisor receives a bonus of 70 points when company profit is low and 190 points when company profit is high.

[10]Buying a winning ball captures employee effort according to the definition of Baiman (1982) because the action is controllable by the employee, creates negative utility for the employee, and results in an increase in the expected profit for the company. Employees can only contribute to company value by increasing their effort while in the real-world, employees can contribute to company value through their effort, knowledge, and ability (Libby & Luft, 1993), I abstract away from how knowledge and ability influence employee contribution by keeping these factors constant. This provides a valid test of the prediction that supervisors reward observable luck.

[11]Using a fixed bonus pool biases against supervisors rewarding observable luck. The size of variable bonus pools is generally determined by objective performance measures such as profit. These objective measures are influenced by luck. Supervisors generate internal anchors that are influenced by the bonus pool size. As a result, supervisors will overweight objective outcomes and, implicitly, luck in their evaluations (Bailey et al., 2011). In contrast, when the size of the bonus pool is fixed, supervisors are probably less likely to use the bonus pool size as an anchor for their bonus and will, therefore, be less likely to reward luck.

Therefore, the supervisor has an incentive to choose a bonus that motivates the employee to buy a high number of winning balls.

After observing the bonus, the employee can confront the supervisor.[12] A confrontation reduces the payoff of both the employee and the supervisor but is more costly to the supervisor (30 points) than to the employee (10 points).[13]

The payoffs functions for each period are:

Employee Payoff = Salary Employee (360 points) - Cost of Buying Winning Balls (between zero and 100 points) + Bonus (between zero and 200 points) - Confrontation costs employee (zero or ten points)

Supervisor = Salary Supervisor (400 points) + Compensation for profit (70 or 190 points) - Confrontation cost supervisor (zero or 30 points)[14]

The supervisor assigns the bonus based on the information presented in the reporting system. The reporting system of the company can produce two measures:

- Profit (high or low). This measure partially captures employee contribution and is influenced by luck.

- Number of winning balls bought. This measure perfectly captures employee contribu-

[12]The confrontation possibility was included to simulate middle-level managers' private incentives when evaluating employees. When employees can confront supervisors, supervisors have private incentives to avoid confrontations because confrontations are costly to the supervisors and are not directly costly to the company. Previous literature has identified middle-level managers' private incentives as one of the main causes why supervisors incorporate employees' self-serving fairness concerns in their evaluations (Arnold & Tafkov, 2019; Bol, 2011; Bol et al., 2016; Deason et al., 2018; Moers, 2005). Given that the integration of employees' self-serving fairness concerns in evaluations is an important part of the theory development, I choose to reproduce a part of middle-level managers' private incentives in my experimental design by allowing employees to confront supervisors.

[13]The fact that confrontations are more costly to the supervisor than to the employee could seem unrealistic. I do not know of any theoretical model that attempts to quantify the costs of confrontations and how they differ between supervisors and employees. Moreover, even if in the real world confrontation costs are higher for employees than for the supervisors, this design choice should not affect the conclusions of this study related to supervisors' evaluations. Lowering the supervisors' cost of confrontation will decrease supervisors' incentives to include employees' self-serving fairness perceptions in evaluations. Given that I find that supervisors incorporate employees' self-serving fairness concerns less than expected, reducing supervisors' incentives to incorporate employees' self-serving fairness concerns should not affect this conclusion.

[14]All monetary amounts are denoted in an experimental currency (points). 200 points have a value of $1.

tion and it is not influenced by luck.

Between subjects, I manipulate the information that supervisors can use when making their bonus decision. The reports presented to the supervisors either contain (1) profit and the number of winning balls bought in the Outcome+Contribution condition or (2) only the number of winning balls bought in the Contribution-Only condition. Supervisors in the Contribution-Only condition learn the profit level for each period after their bonus decision in the last period of the experiment. Therefore, supervisors in the Contribution-Only condition cannot integrate luck in their bonus decisions.

One consideration that guided the parameter choice is worth noting. Supervisors in the real world generally earn more than their employees. This expectation could influence how participants assigned to the supervisor role interpret the experimental instructions (Alekseev et al., 2017). Supervisors could think that they are entitled to always earn more than their employees. Such beliefs could drive supervisors to give lower bonuses to employees when company profit is low. I chose parameters that minimize the impact of this belief on bonuses: supervisors generally earn more than their employees regardless of how much bonus they assign to the employees. For example, in a period in which the employee buys five winning balls and profit is low, the supervisor earns more than the employee even if they assign the maximum bonus (in case the supervisor assigns the maximum bonus, the supervisor earns 470 points and the employee earns 460 points).[15]

Procedure

Participants are recruited from Amazon Mechanical Turk. The task is programmed in OTree (Chen et al., 2016). To facilitate online interactions between MTurk participants, I follow many of the recommendations proposed by Arechar et al. (2018).[16]

[15] Higher differences between the payoff of the employees and the supervisors would likely have been more effective at alleviating this issue. However, I was concerned that bigger differences could cause participants assigned to the employee role to drop out of the study because they found the role allocation unfair.

[16] The task can be played at `http://bit.ly/obs_luck`. The OTree code can be downloaded at `https://bit.ly/obs_luck_code`.

The description of the assignment informs participants about the expected completion time (15-20 minutes) and how much they can earn (between \$1.25 and \$2.95). Participants learn that they will interact with another participant and are asked to start the experiment immediately.[17]

Participants read the instructions and have to correctly answer thirteen understanding questions to participate in the experiment and receive the participation fee.[18] Participants are grouped on the fly: after completing the understanding quiz, participants wait in a lobby for another participant to complete the understanding quiz and join their company. If participants wait for more than eight minutes, they can quit the experiment and receive the participation fee (\$1.25). Participants are randomly assigned to the supervisor or employee role. Participants interact with the same partner for all the eight periods of the experiment. While waiting for the other participant in their company to make their decision, participants see a summary of the instructions and the sequence of events in each period. To minimize potential wealth effects between periods, one period is randomly selected as the payoff period.

After completing the experimental task, participants fill in a post-experimental questionnaire. The questionnaire includes instruments that measure participants' risk aversion (Dohmen et al., 2011) and inequality aversion (Yang et al., 2016). The questionnaire also contains several items and open-ended questions that provide insights into participant's decision-making processes during the experiment.

[17] Asking participants to start immediately is meant to facilitate faster group formation. However, many participants delayed starting the experimental task after accepting the assignment (for example, some participants started the experiment 40 minutes after accepting the assignment). This does not make group formation impossible but it increases the waiting time and the number of participants who are not grouped with a partner because they waited too long. I suspect most MTurk participants likely do not read the announcement page for the HIT because most slots fill out almost immediately after the HIT is posted.

[18] The instructions are still available to participants while answering the understanding checks. If a participant makes a mistake, the computer informs the participant which attention question is answered incorrectly, and wherein the instructions the correct answer can be found. Participants who do not correctly answer all the understanding questions within four attempts are not allowed to participate in the experiment.

4.3.2 Participants

I requested that participants reside in the United States, have completed at least 100 Human Intelligence Tasks (HITs, MTurk's tasks), and have an approval rate of at least 90% on their previous HITs. The experiment could not be opened from a mobile device. In total, data from 258 participants were collected over two days and four sessions. Each session collected data from only one condition to facilitate faster group formation.[19] Participants' age varies from 19 to 73, with a mean of 37.8. 143 participants (55%) were male and 114 participants (44%) were female and one participant did not disclose their gender. Participants reported an average work experience of 16.5 years. 132 participants (51%) indicated that they completed at least a bachelor's degree. Participants received a $1.25 participation fee in addition to the payoff they earned during the experiment. Participants earned an average of $2.37 for an average of twenty minutes of their time. As expected, the payoff of participants in the supervisor role (mean of $2.6) was significantly higher than the payoff of participants in the employee role (mean of $2.1) ($t = 11.64$, $p < 0.01$, two-tailed).

4.4 Results

4.4.1 Randomization Check

I first verify if the random assignment was successful. Participants characteristics did not differ between the two conditions with respect to age, gender, work experience, educations, risk preferences, the propensity to trust, the propensity to reciprocate trust, and inequality aversion. Results of a multiple linear regression indicated that there was no collective significant effect of the measured individual characteristics on being allocated to one of the conditions ($F(8, 249) = 0.64$, $p > 0.10$, $R^2 = 0.02$). Participants also did not differ between the two roles across the same measured characteristics ($F(8, 249) = 0.67$, $p > 0.10$,

[19] Qualifications ensured that a worker could not attempt to complete the experiment more than once or participate in more than one session based on worker IDs. MTurk ensures that MTurk IDs are unique for every individual by asking for the Social Security Number when workers sign up.

$R^2 = 0.02$). These results suggest that random assignment was successful.

Some participants who started the experimental task did not finish the experiment. Due to the nature of conducting research online, some level of attrition is unavoidable. I examine if the attrition was exogenous so that it does not bias the conclusions of this study (Arechar et al., 2018). In my sample, 298 participants started the experimental task and 258 (87%) participants completed the experiment. Twenty participants were excluded during the experiment because they were inactive for two minutes. As a result, the twenty participants that were paired with the inactive participants had to be paid for their time and excluded.[20] I find no evidence that attrition was selective. Dropout rates did not differ between the Outcome+Contribution condition (drop out rate 8.86%) and the ContributionOnly condition (drop out rate 4.29%) ($t = 1.57$, $p > 0.10$, two-tailed). Dropout rates also did not differ between the participants assigned the role of supervisor (drop out rate 8.05%) and participants assigned the role of employee (drop out rate 5.37%) ($t = 0.92$, $p > 0.10$, two-tailed). Given that being assigned the role of supervisor resulted in a higher average payoff, the similar drop-out rate between roles indicates that the dropouts were not caused by lower payoff expectations.

4.4.2 Descriptive Statistics

In total, I collected 1,032 employee-period observations ([258 participants/2 roles] x 8 periods) and the same number of supervisor-period observations.[21] Table 4.2 and Figures 4.1 present the descriptive statistics for the bonuses allocated by the supervisors. Consistent with H1, in the Outcome+Contribution condition, bonuses appear higher when company profit is high than when company profit is low at all levels of employee contribution. Inconsistent with H2, bonuses appear slightly higher in the ContributionOnly condition as compared to

[20] Specifically, participants who were excluded because their partner was inactive received the participation fee plus six cents for each of the eight periods they finished.

[21] All regressions that use multiple observations per participant use robust standard errors clustered at the participant level.

Figure 4.1: Bonus per Contribution for Different Profit Levels

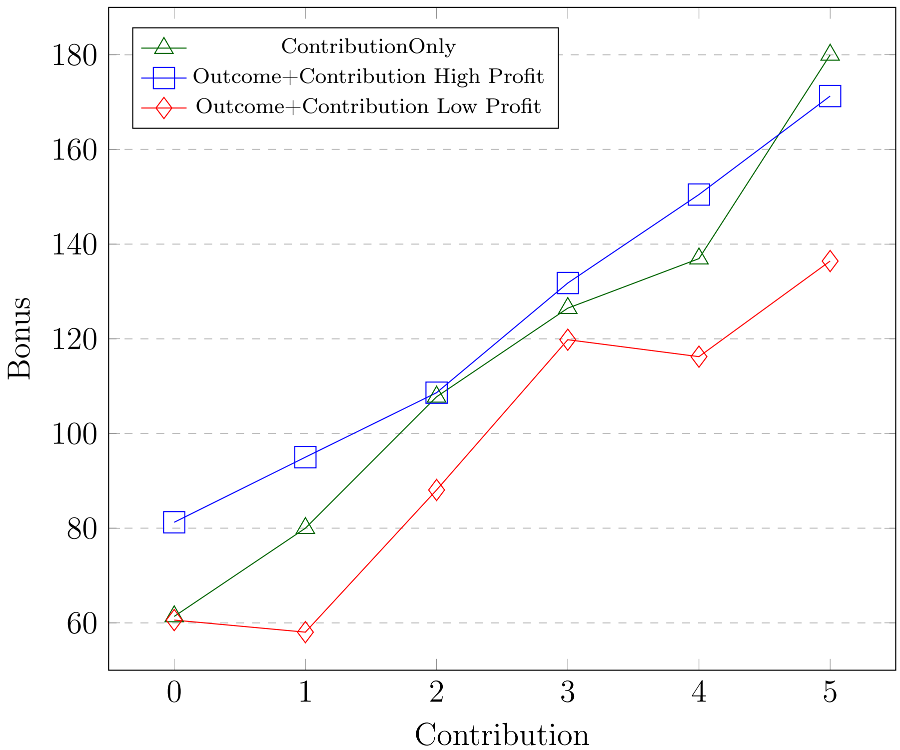


Figure 4.2: Contribution per Period

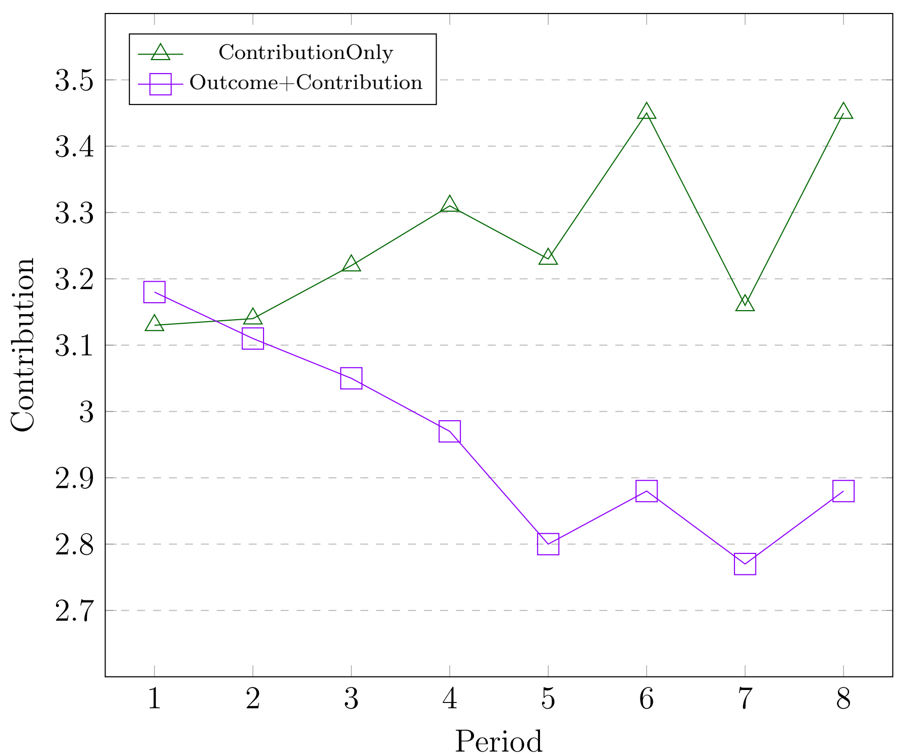


Table 4.2: Bonus per Contribution for Different Profit Levels

Condition		Contribution						
		0	1	2	3	4	5	Overall
ContributionOnly Both Profit Levels	mean	61	80	108	126	137	180	135
	std dev	74	72	76	63	67	55	75
	n	59	14	80	118	63	178	512
Outcome+Contribution Both Profit Levels	mean	64	67	96	125	132	159	118
	std dev	69	61	70	59	71	60	73
	n	86	30	75	104	81	144	520
Outcome+Contribution High Profit	mean	81	95	109	132	150	171	143
	std dev	80	79	76	62	66	51	69
	n	16	7	29	45	38	92	227
Outcome+Contribution Low Profit	mean	61	58	88	120	116	136	98
	std dev	67	54	66	57	71	69	71
	n	70	23	46	59	43	52	293

This table presents the summary statistics for the bonus allocated by the supervisors across the two conditions and across the two profit levels in the Outcome+Contribution condition. In the Outcome+Contribution condition, the bonuses appear higher when profit is high than when profit is low.

Table 4.3: Contribution per Period

Condition		Period								
		1	2	3	4	5	6	7	8	Overall
ContributionOnly	mean	3.13	3.14	3.22	3.31	3.23	3.45	3.16	3.45	3.26
	std dev	1.55	1.58	1.62	1.62	1.67	1.66	1.87	1.74	1.66
Outcome+Contribution	mean	3.18	3.11	3.05	2.97	2.80	2.88	2.77	2.88	2.95
	std dev	1.53	1.63	1.77	1.72	1.89	1.87	1.93	1.82	1.77

This table presents the summary statistics for the employee contribution level across the two conditions and across the eight periods.

the Outcome+Contribution condition at most levels of employee contribution.

Table 4.3 and Figure 4.2 present descriptive statistics for employee contribution. Employee contribution is slightly higher in the ContributionOnly condition (mean = 3.26) as compared to the Outcome+Contribution condition (mean = 2.95). Employees appear to change their contribution across periods differently depending on the experimental condition. Employees in the Outcome+Contribution condition decrease their contribution over time while the employees in the ContributionOnly condition increase their contribution over time.

4.4.3 Hypotheses Tests

Supervisors Reward Observable Luck

H1 predicts that supervisors incorporate observable luck in discretionary evaluations. To test H1, I analyze the bonus data from the Outcome+Contribution condition. I examine the effect of profit and employee contribution on the bonus. After controlling for the effect of *Contribution*, the coefficient of *HighProfit* captures how luck influences the bonus (Brownback & Kuhn, 2019; Stock & Watson, 2015).[22] Therefore, a positive effect of *HighProfit* indicates that supervisors reward observable luck. This is because supervisors can directly observe employee contribution in the Outcome+Contribution condition so, conditional on employee contribution, the profit level only contains information about luck.

Table 4.4, Column 1 reports the results of this regression. Figure 4.1 plots these comparisons. The coefficient of *HighProfit* is significant ($\beta = 26.13$, t = 3.33, $p < 0.01$, two-tailed).[23]

[22] In this setting, luck is the difference between the expected value and the actual value (McKinnon, 2013). For example, if the employee buys five winning balls in a period, the expected value for company profit is 0.7. After the random draw, the value of luck is 0.3 if profit is high and -0.7 if profit is low. I could have captured luck by calculating the residual of the regression of *HighProfit* on *Contribution*. Including this residual in the regression instead of *HighProfit* generates identical coefficients. For ease of exposition, I include *HighProfit* instead of the residual.

[23] The results are qualitatively similar if I generate dummy variables for each contribution level and include these in the regression instead of the linear contribution variable. Thus, the result that supervisors reward observable luck is robust to not assuming that there is a linear relationship between employee contribution and bonus.

This result supports H1.

In the 4.2. Theory section, I argued that supervisors will find it fair to reward luck. Data from the post-experimental questionnaire supports this assumption. I asked participants what should a fair bonus be based on. Participants could choose one of the following options: "how many winning balls were bought", "whether the company profit was high or low", "a combination of the two". I classify participants as having consequentialist fairness views if they thought fair bonuses should be at least partially based on company profit. 43 out of 65 supervisors (66%) reported consequentialist fairness concerns in the Outcome+Contribution condition. Next, I analyze if supervisors with consequentialist fairness concerns are more likely to reward observable luck. Table 4.4, Column 2 reports the results of this regression. The coefficient of the interaction between *HighProfit* and *ConsequentialistViews* is positive and significant (β = 21.23, t = 1.84, p < 0.05, one-tailed) indicating that supervisors with consequentialist views of fairness reward luck more than supervisors with non-consequentialist views of fairness.[24]

Supervisors Asymmetrically Reward Observable Luck

H2 predicts that supervisors reward observable good luck more than they punish observable bad luck. To test H2, I examine the bonus data from the Outcome+Contribution condition.[25] I classify employees as having good or bad luck depending on the combination of employee contribution and the profit level in a given period. Employees have good luck when company profit is high and employee contribution is low (below three). This is because employees generate a high company profit although the employee contribution choice was

[24]The interaction between *HighProfit* and *ConsequentialistViews* becomes marginally significant (β = 16.94, t = 1.40, p < 0.10, one-tailed) if I include dummy variables for each contribution level instead of the linear contribution variable and include the interaction between *Contribution* and *ConsequentialistViews* in the regression (Yzerbyt et al., 2004). Thus, the result that supervisors reward observable luck more when they have consequentialist fairness views is robust to different model specifications.

[25]In the 4.4.4. Robustness Checks section, I perform an additional test for H2 by examining the bonus differences between the Outcome+Contribution and ContributionOnly conditions.

Table 4.4: The Effect of Observable Luck on Bonus

VARIABLES	(1) Bonus	(2) Bonus	(3) Bonus
Contribution	17.06***	16.75***	15.98***
	(3.48)	(3.04)	(4.00)
HighProfit	26.13***	18.03**	16.60
	(7.84)	(8.08)	(18.53)
ConsequentialistViews		-60.17***	
		(13.46)	
HighProfit*ConsequentialistViews		21.23**	
		(11.53)	
HighProfit*Contribution			3.00
			(4.70)
Constant	56.09***	93.63***	58.77***
	(13.17)	(16.18)	(14.06)
Observations	520	520	520

This table presents the results of regressions that analyze supervisors' bonus decisions. The dependent variable is Bonus, which captures how much bonus supervisors allocate to employees in each period. The independent variables are: Contribution, which captures how many winning balls the employee decides to buy (ranging from 0 to 5); HighProfit, which takes the value 1 if the company profit is high and 0 if profit is low; ConsequentialistViews, which takes the value 1 if the participant indicated in the post-experimental questionnaire that they find it fair for profit to influence the bonus and 0 otherwise.
***, **, * denote significance at the 1%, 5%, 10% level (all two-tailed except for the interaction between HighProfit and ConsequentialistViews which is one-tailed). Standard errors (presented in parentheses) are robust and clustered at the supervisor level. H1 predicts a positive effect of HighProfit on Bonus after controlling for Contribution. The coefficient of HighProfit from Column 1 is consistent with H1. The interaction between HighProfit and ConsequentialistViews from Column 2 indicates that supervisors with consequentialist views of fairness reward luck more than supervisors with non-consequentialist views of fairness. H2 predicts a negative interaction effect between Contribution and HighProfit. The coefficient of the interaction between Contribution and HighProfit in Column 3 is inconsistent with H2.

more likely to generate low company profit than high company profit.[26] Similarly, employees have bad luck when company profit is low and employee contribution is high (above three). This is because employees generate low company profit although the employee contribution choice was more likely to result in a high company profit than low company profit. For example, if an employee buys all available winning balls, he/she has bad luck if company profit is low because the employee generates a low profit for the company despite making the choice that is least likely to generate a bad outcome. If I observe that the profit level affects bonuses more when employee contribution is low compared to when employee contribution is high, I can conclude that supervisors reward observable good luck more than they punish observable bad luck. Therefore, a negative interaction effect between *HighProfit* and *Contribution* on *Bonus* would provide support for H2.

Table 4.4, Column 3 reports the results of the regression that tests H2. The interaction between *HighProfit* and *Contribution* is not significant ($\beta = 3.00$, t = 0.64, $p > 0.10$, two-tailed). This result does not support H2.

I further examine the effect of *HighProfit* on *Bonus* for each level of employee contribution. H2 is supported if *HighProfit* affects *Bonus* more when employee contribution is low (below three) than when it is high (above three). Table 4.5 reports the results of these regressions. *HighProfit* has a significant effect on *Bonus* when employee contribution is four ($\beta = 34.24$, t = 2.07, $p < 0.05$, two-tailed) and five ($\beta = 34.85$, t = 3.12, $p < 0.01$, two-tailed). *HighProfit* does not have a significant effect on *Bonus* at any other contribution level (lowest p is 0.20, two-tailed, when employee contribution is one). These results suggest that, contrary to H2, supervisors punish bad luck more than they reward good luck. This result is likely explained by the finding that supervisors punish bad luck when employee contribution is high and do not reward good luck when employee contribution is low.

[26]When employee contribution was lower than three, employees had less than 50% chance of generating a high profit (20%, 30%, 40% respectively)

Table 4.5: The Effect of Observable Luck on Bonus for every level of Employee Contribution

VARIABLES	(1) EC=0 Bonus	(2) EC=1 Bonus	(3) EC=2 Bonus	(4) EC=3 Bonus	(5) EC=4 Bonus	(6) EC=5 Bonus
HighProfit	20.66	36.96	20.56	11.96	34.24**	34.85***
	(26.87)	(27.70)	(21.75)	(15.49)	(16.52)	(11.18)
Constant	60.59***	58.04***	88.07***	119.81***	116.23***	136.42***
	(19.18)	(11.91)	(14.82)	(10.99)	(15.45)	(12.14)
Observations	86	30	75	104	81	144

This table presents the results of regressions that analyze supervisors' bonus decisions. The dependent variable is Bonus, which captures how much bonus supervisors allocate to employees in each period. The independent variable is HighProfit, which takes the value 1 if the company profit is high and 0 if profit is low. Each column presents the effect of HighProfit on Bonus for a different level of employee contribution (EC).

***, **, * denote significance at the 1%, 5%, 10% level (all two-tailed). Standard errors (presented in parentheses) are robust and clustered at the supervisor level. H2 predicts that HighProfit has a higher impact on Bonus when employee contribution is low (below three) than when it is high (above three). HighProfit has a significant effect on Bonus when employee contribution is high (four and five). HighProfit does not have a significant effect on Bonus at any other contribution level. These results suggest that, contrary to H2, supervisors punish bad luck more than they reward good luck.

In the 4.2. Theory section, I argued that supervisors will reward good luck more than they punish bad luck because supervisors anticipate that employees find it fairer to be rewarded for good luck than to be punished for bad luck. The bonus data does not support the conclusion that supervisors reward good luck more than they punish bad luck. However, data from the post-experimental questionnaire supports the assumption that supervisors anticipate that employees find it fairer to be rewarded for good luck than to be punished for bad luck. I asked participants what employees would think a fair bonus should be based on in two situations: when an employee bought one winning ball and profit was high (the employee had good luck) and when an employee bought five winning balls and profit was low (the employee had bad luck). The five options ranged from "only on how many winning balls they bought" to "only on whether the company profit was high or low". I analyze whether participants think employees have asymmetric fairness concerns by calculating the difference between these two answers. A positive difference indicates that respondents consider that employees find it fairer to be rewarded for good luck than punished for bad luck. I find that the difference is significantly different from zero (mean = 0.95, $t = 4.79$, $p < 0.01$, two-tailed), indicating that supervisors anticipate that employees have asymmetric fairness views.

Although supervisors anticipate that employees find it fairer to be rewarded for good luck than to be punished for bad luck, supervisors do not integrate employee self-serving fairness concerns when employee contribution is low. This might be the case because supervisors integrate employees' fairness concerns in evaluations as a form of reciprocity. If this is true, supervisors do not feel the need to reciprocate (offer evaluations that their employees would consider fair) when employee contribution is low. Data from the post-experimental questionnaire supports this conclusion. Supervisors indicated if they thought about allocating a bonus that the employee would consider fair when making the bonus decision. I examine whether the likelihood of supervisors reporting that their bonus decision was influenced by employees' fairness concerns depended on the total employee contribution during the task. Results of an untabulated analysis show that supervisors are more likely to report that their bonus decision was influenced by employees' fairness concerns when the total employee contribution during the task was higher ($\beta = 0.05$, $z = 2.24$, $p < 0.05$, two-tailed). This suggests that supervisors incorporate employee self-serving fairness perceptions in evaluations as a form of reciprocity and that employee contribution has to be high enough to trigger reciprocity in order for employee self-serving fairness perceptions to be incorporated in discretionary evaluations.

Employee Contribution when Supervisors Reward Observable Luck

The RQ asks if employee contribution is lower when supervisors incorporate observable luck in discretionary evaluations. To answer the RQ, I examine employee contribution data from the Outcome+Contribution and ContributionOnly conditions. As seen in the test for H1, supervisors in the Outcome+Contribution condition rewarded observable luck. In contrast, supervisors in the ContributionOnly condition did not reward luck because they did not observe it.[27] Therefore, the ContributionOnly condition allows me to observe employee contribution when supervisors do not reward observable luck and have access to the same amount of information about employee contribution as supervisors in the Outcome+Contribution condition. I regress employee contribution on a *ContributionOnly* dummy. Table 4.6, Column 1 reports the results of this regression. I find no significant difference in employee contribution between the Outcome+Contribution and ContributionOnly conditions ($\beta = 0.31$, $t = 1.22$, $p > 0.10$, two-tailed) indicating that employees do not contribute less when observable luck is incorporated in discretionary evaluations.

Employees likely cannot perfectly anticipate how their supervisors will evaluate them. Instead, employees need to learn their supervisors' evaluation strategy (Choi et al., 2016).

[27]As expected, an untabulated analysis that replicates the regression from Table 4.4, Column 1 with data from the ContributionOnly condition shows an insignificant effect of *HighProfit* after controlling for *Contribution* ($\beta = 1.60$, $t = 0.24$, $p > 0.10$, two-tailed).

Therefore, I also examine if employees change their contribution after they had a chance to learn how their supervisor evaluates them. Table 4.6, Column 2 reports the results of regressing employee contribution on the *ContributionOnly* dummy in the last four periods. The coefficient of *ContributionOnly* is marginally significant (β = 0.49, t = 1.71, p < 0.10, two-tailed) indicating that employee contribution was higher in the last four periods when supervisors did not reward observable luck. Table 4.6, Column 3 reports the results of a regression that examines the different rates of changes in contribution across time between the two conditions. The coefficient of *SecondHalf*, a dummy variable that takes the value 1 in the last four periods of the task, is marginally significant and negative (β = -0.25, t = -1.73, p < 0.10, two-tailed) indicating that employees reduce their contribution across time in the Outcome+Contribution condition, i.e., when they are rewarded for luck. The interaction between *ContributionOnly* and *SecondHalf* is significant and positive (β = 0.37, t = 2.00, p < 0.05, two-tailed) indicating that employee contribution decreases overtime only in the Outcome+Contribution condition. These results suggest that employees decrease their contribution when they learn that supervisors incorporate observable luck in discretionary evaluations.

4.4.4 Robustness Checks

In this section, I examine if the result that supervisors incorporate observable luck in their evaluations (H1) is driven by alternative explanations. Next, I perform an additional test to investigate if supervisors reward observable good luck more than they punish observable bad luck (H2) by examining the bonus differences between the Outcome+Contribution and ContributionOnly conditions.

Incomplete Understanding of Consequences

One alternative explanation for the H1 result is that supervisors do not completely think through the consequence of their actions and therefore do not understand that rewarding only employee contribution instead of performance (contribution and luck) can better motivate

Table 4.6: The Effect of Integrating Observable Luck on Employee Contribution

VARIABLES	(1) Contribution	(2) Contribution	(3) Contribution
ContributionOnly	0.31	0.49*	0.12
	(0.25)	(0.29)	(0.25)
SecondHalf			-0.25*
			(0.14)
ContributionOnly*SecondHalf			0.37**
			(0.19)
Constant	2.98***	2.83***	3.08***
	(0.18)	(0.22)	(0.18)
Observations	1,032	516	1,032

This table presents the results of regressions that analyze employees' contributions. The dependent variable is Contribution, which captures how many winning balls the employee decides to buy in each period (ranging from 0 to 5). The independent variables are: ContributionOnly, which takes the value 1 if the employee was assigned to the ContributionOnly condition and 0 if the employee was assigned to the Outcome+Contribution condition; SecondHalf, which takes the value 1 in the last four periods of the task and 0 otherwise.
***, **, * denote significance at the 1%, 5%, 10% level (all two-tailed). Standard errors (presented in parentheses) are robust and clustered at the employee level. Column 1 reports the result of the regression for all periods and Column 2 reports the results of the regression in the last four periods. The RQ asks if the ContributionOnly variable will affect Contribution. The coefficient of ContributionOnly from Column 1 suggests that employees do not change their contribution when supervisors reward observable luck. The coefficient of ContributionOnly from Column 2 suggests that employee contribution decreases after employees learn that supervisors reward observable luck. The negative coefficient SecondHalf and the positive interaction between ContributionOnly and SecondHalf from Column 3 indicate that employee contribution decreases overtime only when supervisors reward observable luck.

employees to increase their contribution. Consistent with this explanation, Krishnan et al. (2005) find that people make compensation decisions that are inconsistent with agency theory because they use incomplete mental models. Results from the post-experimental questionnaire suggest that this explanation cannot fully explain why supervisors reward observable luck.

In the post-experimental questionnaire, supervisors had to indicate what type of bonus would best motivate the employee to buy as many winning balls as possible. Supervisors could choose one of the following options: "how many winning balls the employee bought", "whether the company profit was high or low" or "a combination of the two". I classify supervisors as having beliefs consistent with traditional agency theory if they believed that employees are best motivated by rewarding them based only on the number of winning balls bought.[28] In the Outcome+Contribution condition, 24 out of 65 supervisors (37%) have beliefs consistent with agency theory. In an untabulated analysis, I examine the effect of *Contribution*, *HighProfit*, *NonAgencyBeliefs* (takes the value 0 if the supervisors reported agency consistent beliefs in the post-experimental questionnaire and 1 otherwise) and the interaction between *HighProfit* and *NonAgencyBeliefs* on *Bonus* in the Outcome+Contribution condition. The coefficient of the interaction between *HighProfit* and *NonAgencyBeliefs* is not significant ($\beta = 13.39$, t = 1.07, p > 0.10, one-tailed) indicating that supervisors with beliefs different to agency theory about how to motivate employees reward luck to a similar extent as supervisors with beliefs consistent with agency theory. Moreover, the coefficient of *HighProfit* is significant and positive ($\beta = 23.40$, t = 2.57, p < 0.05, two-tailed) indicating that supervisors with beliefs consistent with agency theory reward observable luck. These results do not support the logic that supervisors reward observable luck because they do not

[28]Note that supervisors who had views inconsistent with agency theory did not necessarily use an incomplete mental model like participants in Krishnan et al. (2005). These participants could have believed that rewarding luck increases employee motivation because evaluations that incorporate luck are perceived as fairer by employees (Fehr et al., 2009).

understand that rewarding luck could decrease employee motivation.

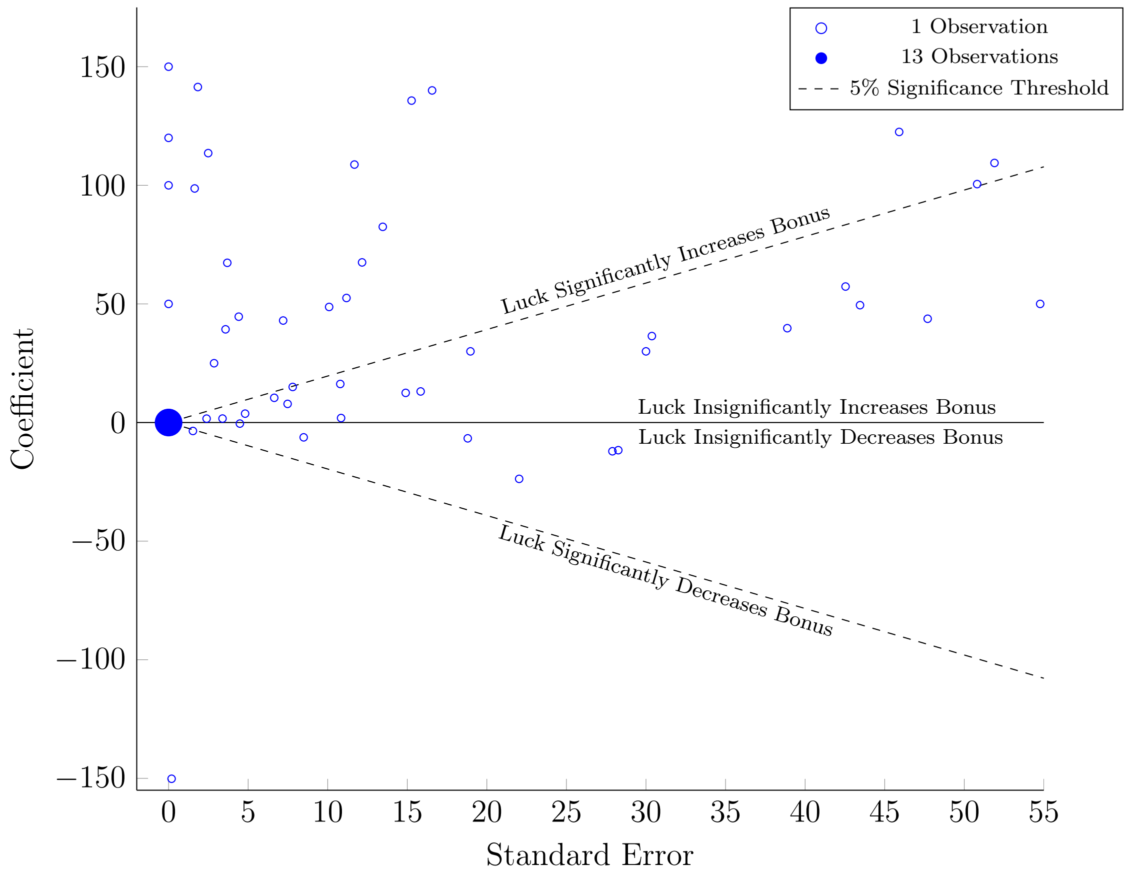
Outliers

The H1 result may be driven entirely by a few supervisors who completely ignore contribution and only reward company profit. For example, a few supervisors could be assigning the full bonus when profit is high and no bonus when profit is low regardless of the employee's contribution level. If this would be the case, the H1 result could be driven by outliers. To examine this possibility, I calculate supervisor-specific estimates of the effect of observable luck on bonuses by replicating the main test for H1 (the regression model from Table 4.4, Column 1) for each supervisor separately (Brownback & Kuhn, 2019). Figure 4.3 plots these estimates. Individual estimates are widespread indicating that the H1 result is not driven by outliers.

Confusion

Despite needing to correctly answer all the understanding questions, supervisors could have been confused about the experimental instructions. Supervisors may have rewarded luck because they did not understand how employee contribution affects their payoff. I investigate if results are consistent with this alternative explanation. All participants needed to correctly answer all thirteen understanding questions within four attempts in order to participate in the experiment. If a participant made a mistake, the computer informed the participant which attention question was answered incorrectly, and wherein the instructions the correct answer could be found. It is possible that participants used this feedback to guess the correct answer to the multiple-choice attention questions through a process of elimination. If this is true, participants who correctly answered all attention questions from the first attempt have a better understanding of the experimental instructions as compared to all other participants. In the Outcome+Contribution condition, 26 out of 65 supervisors (40%) correctly answered all the attention questions from the first attempt. I replicate the main test for H1 (the regression model from Table 4.4, Column 1) with the subsample of supervisors

Figure 4.3: Individual Estimates of the Effect of Observable Luck on Bonus



This graphic presents individual estimates of the effect of luck on bonuses obtained by running separate regressions for each supervisor in the Outcome+Contribution condition. The dependent variable is Bonus, which captures how much bonus supervisors allocate to employees in each period. The independent variables are: Contribution, which captures how many winning balls the employee decides to buy (ranging from 0 to 5) and HighProfit, which takes the value 1 if the company profit is high and 0 if profit is low. Each estimate was based on eight observations. Standard errors are robust. Out of the 65 supervisors in the Outcome+Contribution condition, four were dropped from the analysis because they did not have any variation in the profit level across the eight periods. Individual estimates are widespread indicating that the H1 result is not driven by outliers. The coefficient of HighProfit was statistically significant at a 5% level (two-tailed) for 22 out of 61 supervisors. HighProfit was dropped out of the regression entirely for thirteen supervisors because it had no explanatory power.

who correctly answered all the attention questions on their first attempt. An untabulated analysis finds qualitatively similar results as the main test of H1 (coefficient of *HighProfit* is $\beta = 33.15$, t = 3.49, p < 0.01, two-tailed). Therefore, to the extent that answering all thirteen attention questions on the first attempt indicates a better understanding of the instructions, this result does not support the logic that supervisors reward observable luck because they do not understand the instructions.

Low-Quality Responses

Some MTurk participants provide low-quality responses (Dennis et al., 2020). Only these participants may be rewarding observable luck. Dennis et al. (2020) recommend using the answers to the open-ended post-experimental questions to assess the quality of MTurk responses. I, therefore, examine if the length of supervisors' response to the optional open-ended question moderates the propensity to reward observable luck. I still find evidence consistent with H1 if I examine the subsample of supervisors with responses longer than 88 characters (the median) (coefficient of *HighProfit* is $\beta = 33.20$, t = 3.11, p < 0.01, two-tailed). Moreover, if I interact the length of the open-ended response with *HighProfit*, the interaction is not statistically significant ($\beta = 0.04$, t = 0.69, p > 0.10, two-tailed). These results do not support the logic that supervisors reward observable luck because MTurk workers provide low-quality responses.

Supervisors Asymmetrically Reward Luck - Additional Evidence

I perform an additional test to investigate if supervisors reward observable good luck more than they punish observable bad luck (H2) by examining the bonus differences between the Outcome+Contribution and ContributionOnly conditions. Because supervisors in the ContributionOnly condition did not observe luck, they could not reward luck. Therefore, the ContributionOnly condition allows me to observe how supervisors allocate bonuses when they have access to the same amount of information about employee contribution as supervisors in the Outcome+Contribution condition but do not reward luck. I compare the differences in bonuses between the (1) ContributionOnly condition and the Outcome+Contribution condition when profit is low to (2) the differences in bonus between the ContributionOnly condition to the bonus in the Outcome+Contribution condition when profit is high. If supervisors reward good and bad luck symmetrically, I expect the bonus differences between the ContributionOnly condition and the Outcome+Contribution condition when profit is high to be similar to the bonus differences between the ContributionOnly condition and the Outcome+Contribution condition when profit is low. If supervisors reward good luck more than they punish bad luck, I expect the bonus differences between the ContributionOnly condition and the Outcome+Contribution condition when profit is high (when the employee was lucky) to be higher than the bonus differences between the ContributionOnly condition and the Outcome+Contribution condition when profit is low (when the employee was unlucky).

Figure 4.1 plots these comparisons. The coefficient of *ContributionOnly* from Table 4.7, Column 1 compares the bonuses in the ContributionOnly condition to the bonuses in the Outcome+Contribution conditions when profit is high. The coefficient of *ContributionOnly* is not significant (β = -1.60, t = -0.16, p $>$ 0.10, one-tailed) indicating that, after controlling for *Contribution*, the bonuses are similar between the ContributionOnly condition and the Outcome+Contribution condition when profit is high. Table 4.7, Column 2 compares the bonuses in the ContributionOnly condition to the bonuses in the Outcome+Contribution conditions when profit is low. The coefficient of *ContributionOnly* is significant (β = 20.25, t = 1.90, p $<$ 0.05, one-tailed) indicating that, after controlling for employee contribution, the bonuses are higher in the ContributionOnly condition as compared to the Outcome+Contribution condition when profit is low. Therefore, the bonus differences between the ContributionOnly condition and the Outcome+Contribution condition when profit is high are lower than the bonus differences between the ContributionOnly condition and the Outcome+Contribution condition when profit is low. These results do not support H2 and indicate that supervisors

Table 4.7: The Asymmetric Effect of Observable Luck on Bonus

VARIABLES	(1) HighProfit Bonus	(2) LowProfit Bonus
Contribution	22.19***	20.44***
	(2.84)	(2.68)
ContributionOnly	-1.60	20.25**
	(9.99)	(10.58)
Constant	63.87***	47.75***
	(13.30)	(10.95)
Observations	739	805

This table presents the results of regressions that analyze supervisors' bonus decisions. The dependent variable is Bonus, which captures how much bonus supervisors allocate to employees. The independent variables are: Contribution, which captures how many winning balls the employee decides to buy (ranging from 0 to 5) and ContributionOnly, which takes the value 1 if the employee was assigned to the ContributionOnly condition and 0 if the employee was assigned to the Outcome+Contribution condition.
***, **, * denote significance at the 1%, 5%, 10% level (all two-tailed except for ContributionOnly which is one-tailed). Standard errors (presented in parentheses) are robust and clustered at the supervisor level. Column 1 reports the result of a regression that contains only observations when profit was high in the Outcome+Contribution condition and all observations from the ContributionOnly condition. Column 2 reports a result of a regression that contains only observations when profit was low in the Outcome+Contribution condition and all observations from the ContributionOnly condition. An additional test for H2 predicts a higher coefficient (in absolute terms) of ContributionOnly in Column 1 than in Column 2. These results do not support H2. The results suggest that supervisors punish employees for bad luck more than they reward employees for good luck.

punish employees for bad luck more than they reward employees for good luck.

4.4.5 Additional Analyses

Consequentialist Fairness Views and Employee Behavior

I find that 82 out of 129 employees (64%) have consequentialist views of fairness, that is, they find it fair to be rewarded for luck.[29] I next examine if employees with consequentialist fairness views react differently than employees with non-consequentialist views to evaluations

[29]Recall that I categorize participants as having consequentialist fairness views when they indicate in the post-experimental questionnaire that a fair bonus should be at least partially based on luck.

that incorporate luck. First, I examine if employees with consequentialist fairness views regard evaluations as fairer when they perceive evaluations are more influenced by luck. Perceived evaluation fairness is an important determinant of employee motivation (Asay et al., 2019; Bol, 2011; Colquitt et al., 2001; Giraud et al., 2008; Voußem et al., 2016).[30] Although previous literature argues that employees find it unfair to be rewarded for luck (Bol, 2011; Bol et al., 2015; Voußem et al., 2016), employees with consequentialist fairness views may find it fairer to be partially rewarded for luck. I measured the perceived weight of luck in evaluations by asking employees in the post-experimental questionnaire to indicate how much they agree with the following sentence: "luck played a big part in how much bonus I got". I measured the perceived fairness of the evaluation by asking employees to indicate how much they agree with the following sentence: "the bonuses I received were fair". Both were measured using a five-point Likert scale. I regress the perceived fairness of the evaluation on the perceived weight of luck in evaluations for employees with consequentialist and non-consequentialist fairness views in the Outcome+Contribution condition. I find that a higher perceived weight of luck increases the perceived fairness of the evaluation when employees have consequentialist fairness views (β = 0.37, t = 2.30, $p < 0.05$, two-tailed). When employees have non-consequentialist fairness views, a higher perceived weight of luck does not significantly affect the perceived fairness of the evaluation (β = -0.13, t = -0.51, $p > 0.10$, two-tailed).

Second, I examine if contribution levels differ between employees with consequentialist fairness views and employees with non-consequentialist fairness views when supervisors can reward luck. On the one hand, having consequentialist fairness views could be associated with a higher contribution level because these employees can receive evaluations

[30]Consistent with the idea that fairness concerns are important for employee motivation, I find that employees' fairness perceptions are positively associated with employee contribution (β = 0.55, t = 6.22, $p < 0.01$, two-tailed) (Bol, 2011; Fehr et al., 2009) and a negative associated with confrontations (β = -0.57, z = -4.00, $p < 0.01$, two-tailed) (Bol et al., 2016).

they perceive as fairer when supervisors can reward luck (Bol, 2011; Fehr et al., 2009). On the other hand, having consequentialist fairness views could be associated with a lower contribution level because such employees could find it fair to receive a maximum bonus by having good luck instead of maximizing their contribution levels. In contrast, employees with non-consequentialist fairness views likely find it fair to only receive the maximum bonus when they maximize their contribution. I regress *Contribution* on *ConsequentialistViews* in the Outcome+Contribution condition. An untabulated analysis shows that employees with consequentialist fairness views have a lower contribution level compared to employees with non-consequentialist fairness views when they can be rewarded for luck (β = -1.05, t = -2.29, p $<$ 0.05, two-tailed). Employees with consequentialist fairness views are less likely to maximize their contribution by buying five winning balls in all eight periods (β = -3.62, z = -3.21, p $<$ 0.01, two-tailed). Out of the eight employees in the Outcome+Contribution condition who maximized their contribution, only one employee reported consequentialist fairness views. If I only examine employees who did not maximize their contribution in all eight periods (57 out of 65 employees), the effect of consequentialist fairness views on contribution becomes insignificant (β = -0.12, t = -0.20, p $>$ 0.10, two-tailed). These results suggest that employees with consequentialist fairness views are less likely than employees with non-consequentialist fairness views to maximize their contribution.

Third, I investigate if employees with consequentialist fairness views confront supervisors who do not reward luck because they find evaluations that do not incorporate luck as unfair (Bol et al., 2016). I examine the confrontation decisions of the subpopulation of employees with consequentialist views of fairness. I regress *Confrontation* on *Contribution*, *HighProfit* and *Bonus*. Similar to the test for H1, after controlling for the effect of *Contribution*, the coefficient of *HighProfit* captures how luck influences confrontations. I include *Bonus* to increase the precision of the test by reducing the error variance (Wooldridge, 2016). Table 4.8, Column 1 reports the results of this regression within the Outcome+Contribution condition. The coefficient of *HighProfit* is not significant ($\beta = 0.29$, z = 0.85, p > 0.10, one-tailed) indicating that employees do not confront supervisors more when profit is high in the Outcome+Contribution condition. However, this could be the case because supervisors already incorporated luck in their evaluation in anticipation of employees' fairness concerns. To alleviate this issue, I examine confrontations in the ContributionOnly conditions where supervisors cannot integrate luck in evaluations because they do not observe it. Table 4.8, Column 2 reports the results of this regression. The coefficient of *HighProfit* is significant ($\beta = 0.99$, z = 1.8, p < 0.05, one-tailed) indicating that employees confront supervisors more when they have good luck than when they have bad luck even if supervisors cannot observe and react to their luck. As expected, employees in the ContributionOnly with non-consequentialist fairness views do not confront supervisors more when they have good luck than when they have bad luck (β = -0.45, z = -1.24, p > 0.10, one-tailed). These results suggest that employees who find it fair to be rewarded for luck confront supervisors for not rewarding luck.

Inequality Aversion and Supervisors' Evaluations

I examine whether supervisors' inequality aversion influences the degree to which they reward observable luck. Inequality aversion and fairness are closely related concepts (Brosnan & de Waal, 2014). It is possible that supervisors reward observable luck because they find it fair to reduce the inequality between themselves and their employees. Therefore, supervisors who are more concerned about reducing inequality should reward observable luck more. In the post-experimental questionnaire, I measure participants' inequality aversion using six hypothetical items from the scale developed by Yang et al. (2016). Each item asks participants to choose between increasing their own payoff and decreasing the inequality between themselves and another participant. I measure *InequalityAversion* as the total number of choices made by participants in which they sacrificed some of their own payoff to reduce inequality. I test whether there is a significant interaction effect between *Inequal-*

ityAversion and *HighProfit* on *Bonus* in the Outcome+Contribution condition. I find that the interaction effect is significant (β = 12.61, t = 2.65, p < 0.05, two-tailed) indicating that supervisors reward observable luck more when they are more inequality averse. Some people have asymmetric inequality aversion in the sense that they avoid envy (disutility from earning less than others) more than guilt (disutility from earning more than others). Three of the six items that measure inequality aversion measure quilt and three measure envy. In two regressions, I find that the interaction between *Envy* and *HighProfit* is significant (β =

Table 4.8: The Effect of Observable Luck on Employee Confrontation

	(1)	(2)
	Outcome+Contribution	ContributionOnly
VARIABLES	Confront	Confront
Contribution	0.34***	0.79***
	(0.13)	(0.22)
HighProfit	0.29	0.99**
	(0.34)	(0.55)
Bonus	-0.02***	-0.03***
	(0.01)	(0.01)
Constant	-0.83**	-1.48***
	(0.40)	(0.57)
Observations	392	264

This table presents the results of logit regressions that analyze employees' confrontations. Only employees that indicated in the post-experimental questionnaire that they find it fair for profit to influence the bonus are included in these regressions. The dependent variable is Confront, which takes the value 1 if the employee confronted the supervisor in a period and 0 otherwise. The independent variables are: Contribution, which captures how many winning balls the employee decides to buy (ranging from 0 to 5); HighProfit, which takes the value 1 if the company profit is high and 0 if profit is low; Bonus, which captures how much bonus supervisors allocate to employees. ***, **, * denote significance at the 1%, 5%, 10% level (all two-tailed except for HighProfit which is one-tailed). Standard errors (presented in parentheses) are robust and clustered at the employee level. Column 1 reports the result in the Outcome+Contribution condition and Column 2 reports the results in the ContributionOnly condition. The coefficient of HighProfit is insignificant in the Outcome+Contribution condition and significant in the ContributionOnly condition. These results suggest that employees who find it fair to be rewarded for luck confront supervisors who do not reward observable luck.

16.22, $t = 1.99$, $p < 0.05$, two-tailed) and the interaction *Guilt* and *HighProfit* is insignificant ($\beta = 8.08$, $t = 1.36$, $p > 0.10$, two-tailed). These results suggest that supervisors reward observable luck more when they are more concerned about earning less than others.

Employees' Fairness Views about Rewarding Good Luck and Punishing Bad Luck

Finally, I present evidence from the post-experimental questionnaire about whether employees find if fair for good luck to be rewarded and bad luck to be punished. Recall that I asked participants to indicate what employees would think a fair bonus should be based on in two situations: when an employee bought one winning ball and profit was high (the employee had good luck) and when an employee bought five winning balls and profit was low (the employee had bad luck). Out of the 65 employees in the Outcome+Contribution Condition, more employees (56, 86%) thought the bonus should be at least partially based on luck when the employee had good luck than when the employee had bad luck (43, 66%) (mean difference = 0.20, $t = 4.00$, $p < 0.01$, two-tailed). This result suggests that employees have asymmetric fairness views and that a significant portion of the employees find it fair for bad luck to be at least partially incorporated in discretionary evaluations.

4.5 Discussion and Conclusion

This study uses an experiment to examine if middle-level supervisors reward observable luck in their evaluation decisions and the effect this has on employee behavior. I find that many supervisors do not behave according to the controllability principle and partially reward luck even if the effect of luck on performance is perfectly observable. Many supervisors and employees report that they find evaluations that incorporate luck fairer than evaluations that are strictly based on employee contribution. Supervisors reward good and bad luck asymmetrically. Contrary to my initial prediction, supervisors punish bad luck more than they reward good luck. Employees' contribution is lower when supervisors reward observable luck but only after employees learn how supervisors evaluate them through repeated interactions.

I contribute to the discretionary evaluation literature in several ways. First, I contribute to the literature that examines how supervisors' fairness concerns or cognitive limitations can decrease the theoretical contracting benefits of discretionary evaluations (Bailey et al., 2011; Baiman & Rajan, 1995; Krishnan et al., 2005; Lipe & Salterio, 2000; Rajan & Reichelstein, 2006). Because supervisors reward observable luck, allowing supervisor discretion may not always lower the impact of luck on employee compensation even if supervisors have access to additional non-contractible information.

Moreover, the finding that supervisors reward observable luck can improve our understanding of a puzzling empirical phenomenon: although we have good reasons to believe that rewarding luck can be detrimental to employee motivation (Bol, 2008), supervisors seem to routinely disregard the controllability principle when evaluating employees (Bol et al., 2015; Giraud et al., 2008; Merchant, 1987). Previous literature argues that supervisors reward luck because it promotes adaptive behavior (Bol et al., 2015; Simons, 2010) and because it is difficult to objectively determine how luck influences performance (Giraud et al., 2008; Merchant, 1987). Both of these explanations imply that supervisors reward luck because they have insufficient information about how luck affects performance. By designing an experiment in which supervisors can perfectly observe how luck affects performance, I can provide evidence that a previously unknown factor causes supervisors to reward luck. Specifically, I propose and find evidence that fairness concerns cause supervisors to reward luck.

Second, I expand our understanding of how fairness concerns influence discretionary evaluations. Accounting researchers have argued that a sense of fairness causes supervisors to only hold employees accountable for factors that employees can immediately control (Arnold & Tafkov, 2019; Bol, 2011; Chan, 2018; Maas et al., 2012). In this paper, I add to our understanding of how controllability influences fairness perception by showing that some supervisors and employees consider it fair for completely uncontrollable factors to be incorporated into evaluations.

Third, I contribute to the literature that examines how employees' self-serving fairness perceptions influence discretionary evaluations (Arnold et al., 2018; Arnold & Tafkov, 2019; Bol et al., 2016). Previous literature argues that employees' self-serving fairness perceptions cause leniency in discretionary evaluations partially because supervisors want to avoid confrontations with their employees (Arnold & Tafkov, 2019; Bol, 2011; Bol et al., 2016; Deason et al., 2018; Moers, 2005). I find that although supervisors anticipate employees' self-serving fairness perceptions and confrontations with the employees are costly, supervisors do not indiscriminately integrate employees' fairness perceptions in discretionary evaluations. Specifically, supervisors do not integrate these self-serving fairness perceptions when employees have a low contribution to company value. This suggests that supervisors incorporate employee self-serving fairness perceptions as a form of reciprocity towards employees who already have a high enough contribution to trigger reciprocity. This finding can help explain why leniency does not always decrease employee contribution as would be predicted by agency theory (Bol, 2011). Supervisors could be using leniency as a motivational mechanism if employees understand that supervisors are only lenient toward employees who have a high enough contribution to company value.

Fourth, I examine how employees change their contribution and confrontational behavior when supervisors reward observable luck. Although some employees find it fair to be rewarded for luck, rewarding observable luck results in a lower employee contribution after employees learn how supervisors evaluate them. Regarding confrontational behavior, I find evidence that employees with consequentialist views of fairness sacrifice part of their own payoff to confront supervisors for not rewarding luck. These results suggest that when supervisors decide whether to filter out luck from evaluation, they are faced with a trade-off. On the one hand, supervisors should filter out luck because employees decrease their contribution when they learn that supervisors reward luck. On the other hand, supervisors should reward employees who have good luck in order to avoid confrontations with employees that have consequentialist fairness views. Supervisors who initially base evaluations exclusively on employee contribution could, through repeated interactions with these employees, integrate luck into their evaluations to avoid costly confrontations. Practitioners or researchers who design interventions that aim to increase the weight of employee contributions in discretionary evaluations (Berger et al., 2013; Bol et al., 2018, 2016; Demeré et al., 2019) should therefore also consider whether employees find it fair to be rewarded for luck and what control mechanisms could change employee fairness perceptions towards less consequentialist fairness views.

Future research can build on the limitations of my study in several ways. First, although I provide evidence that supervisors reward observable luck and that such behavior can decrease the motivational effect of discretionary evaluations, I do not examine how organizations can motivate supervisors to ignore luck in their evaluations. Future research can investigate mechanisms that decrease the weight of observable luck in discretionary evaluations. For example, a possible cheap intervention is to attempt to nudge supervisors and employees into accepting that evaluations should be based on employee contribution by communicating that the company values such evaluations. Such informal controls seem to have a high impact on how employees judge trade-offs (Kachelmeier et al., 2016). Second, in my study, employees' ability and knowledge are held constant. Employees likely have less control over their knowledge and ability as compared to their control over their effort (Chan, 2018). Future research can investigate how these differences in controllability affect fairness perceptions regarding discretionary evaluations. For example, a supervisor with a consequentialist view of fairness should find it fair to give high evaluations to low-effort, high-ability employees. However, such an evaluation strategy will not motivate the employee to increase their effort and could send the wrong message in the organization (Choi et al., 2016). Third, the current study uses a situational question to classify participants' fairness concerns. A situation-independent scale could help us better understand how fairness influences discretionary evaluations.

Bibliography

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.

Akbaş, M., Ariely, D., & Yuksel, S. (2019). When is inequality fair? An experiment on the effect of procedural justice and agency. *Journal of Economic Behavior & Organization*, *161*, 114–127.

Alekseev, A., Charness, G., & Gneezy, U. (2017). Experimental methods: When and why contextual instructions are important. *Journal of Economic Behavior & Organization*, *134*, 48–59.

Amabile, T. M., & Glazebrook, A. H. (1982). A negativity bias in interpersonal evaluation. *Journal of Experimental Social Psychology*, *18*(1), 1–22.

Anand, V., Webb, A., & Wong, C. (2019). Mitigating the potentially demotivating effects of early and frequent feedback about goal progress. *Working Paper*.

Arechar, A. A., Gächter, S., & Molleman, L. (2018). Conducting interactive experiments online. *Experimental Economics*, *21*(1), 99–131.

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.

Arnold, M. C., & Tafkov, I. D. (2019). Managerial discretion and task interdependence in teams. *Contemporary Accounting Research*, *36*(4), 2467–2493.

Arrow, K. J. (1969). Classificatory notes on the production and transmission of technical knowledge. *The American Economic Review*, *59*(2), 29–35.

Arya, A., & Glover, J. C. (2014). On the upsides of aggregation. *Journal of Management Accounting Research*, *26*(2), 151–166.

Asay, H. S., Garrett, J., & Tayler, W. B. (2019). Fairness and controlling for the uncontrollable. *Working Paper, Available at SSRN 2823976*.

Azoulay, P., Graff Zivin, J. S., & Manso, G. (2011). Incentives and creativity: Evidence from the academic life sciences. *The RAND Journal of Economics*, *42*(3), 527–554.

Babcock, L., Wang, X., & Loewenstein, G. (1996). Choosing the wrong pond: Social comparisons in negotiations that reflect a self-serving bias. *The Quarterly Journal of Economics*, *111*(1), 1–19.

Bailey, W. J., Hecht, G., & Towry, K. L. (2011). Dividing the pie: The influence of managerial discretion extent on bonus pool allocation. *Contemporary Accounting Research*, *28*(5), 1562–1584.

Baiman, S. (1982). Agency research in management accounting: A survey. *Journal of Accounting Literature*, *1*(1), 154–210.

Baiman, S., & Rajan, M. V. (1995). The informational advantages of discretionary bonus schemes. *The Accounting Review*, 557–579.

Baker, G., Gibbons, R., & Murphy, K. J. (1994). Subjective performance measures in optimal incentive contracts. *The Quarterly Journal of Economics*, *109*(4), 1125–1156.

Baker, G., & Jorgensen, B. (2003). Volatility, noise and incentives. *Working Paper*.

Baron, J., & Hershey, J. C. (1988). Outcome bias in decision evaluation. *Journal of Personality and Social Psychology*, *54*(4), 569.

Baumeister, R. F., Bratslavsky, E., Finkenauer, C., & Vohs, K. D. (2001). Bad is stronger than good. *Review of General Psychology*, *5*(4), 323–370.

Benndorf, V., Rau, H. A., & Sölch, C. (2018). Minimizing learning behavior in repeated real-effort tasks. *Working Paper, Available at SSRN 2503029*.

Bentley, J., & Stubbs, K. (2020). The effect of performance information aggregation on bias in subjective performance evaluation. *Working Paper*.

Berger, J., Harbring, C., & Sliwka, D. (2013). Performance appraisals and the impact of forced distribution—an experimental investigation. *Management Science*, *59*(1), 54–68.

Berger, L. (2019). The importance of task complexity when rewarding nonfinancial performance. *Journal of Management Accounting Research*, *31*(2), 1–17.

Bertrand, M., & Mullainathan, S. (2001). Are CEOs rewarded for luck? the ones without principals are. *The Quarterly Journal of Economics*, *116*(3), 901–932.

Bhargava, S., & Fisman, R. (2014). Contrast effects in sequential decisions: Evidence from speed dating. *Review of Economics and Statistics*, *96*(3), 444–457.

Bohn, R. E. (1995). Noise and learning in semiconductor manufacturing. *Management Science*, *41*(1), 31–42.

Bol, J. C. (2008). Subjectivity in compensation contracting. *Journal of Accounting Literature*, *27*(1), 1–24.

Bol, J. C. (2011). The determinants and performance effects of managers' performance evaluation biases. *The Accounting Review*, *86*(5), 1549–1575.

Bol, J. C., Aguiar, A. B., Lill, J. B., & Coelho, A. C. (2018). Peer-level calibration committees. *Working Paper, Available at SSRN 3161692*.

Bol, J. C., Hecht, G., & Smith, S. D. (2015). Managers' discretionary adjustments: The influence of uncontrollable events and compensation interdependence. *Contemporary Accounting Research*, *32*(1), 139–159.

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.

Bol, J. C., & Smith, S. D. (2011). Spillover effects in subjective performance evaluation: Bias and the asymmetric influence of controllability. *The Accounting Review*, *86*(4), 1213–1230.

Bonner, S. E., & Sprinkle, G. B. (2002). The effects of monetary incentives on effort and task performance: Theories, evidence, and a framework for research. *Accounting, Organizations and Society*, *27*(4-5), 303–345.

Brazel, J. F., Jackson, S. B., Schaefer, T. J., & Stewart, B. W. (2016). The outcome effect and professional skepticism. *The Accounting Review*, *91*(6), 1577–1599.

Brosnan, S. F., & de Waal, F. B. (2014). Evolution of responses to (un) fairness. *Science*, *346*(6207), 314–321.

Brownback, A., & Kuhn, M. A. (2019). Understanding outcome bias. *Games and Economic Behavior*, *117*, 342–360.

Cadsby, C. B., Song, F., & Tapon, F. (2016). The impact of risk-aversion and stress on the incentive effect of performance-pay. In *Experiments in Organizational Economics* (pp. 189–227). Emerald Group Publishing Limited.

Cadsby, C. B., Song, F., & Zubanov, N. (2019). The "sales agent" problem: Effort/leisure allocation under performance pay as behavior towards risk. *Economic Inquiry*, *57*(4), 1997–2016.

Camerer, C. F., Ho, T.-H., & Chong, J.-K. (2004). A cognitive hierarchy model of games. *The Quarterly Journal of Economics*, *119*(3), 861–898.

Camerer, C. F., Ho, T.-H., & Chong, J. K. (2015). A psychological approach to strategic thinking in games. *Current Opinion in Behavioral Sciences*, *3*, 157–162.

Campbell, D. (2008). Nonfinancial performance measures and promotion-based incentives. *Journal of Accounting Research*, *46*(2), 297–332.

Campbell, D., Epstein, M. J., & Martinez-Jerez, F. (2011). The learning effects of monitoring. *The Accounting Review*, *86*(6), 1909–1934.

Cappelen, A. W., Hole, A. D., Sørensen, E. Ø., & Tungodden, B. (2007). The pluralism of fairness ideals: An experimental approach. *American Economic Review*, *97*(3), 818–827.

Cappelen, A. W., Konow, J., Sørensen, E. Ø., & Tungodden, B. (2013). Just luck: An experimental study of risk-taking and fairness. *American Economic Review*, *103*(4), 1398–1413.

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.

Casas-Arce, P., Lourenço, S. M., & Martínez-Jerez, F. A. (2017). The performance effect of feedback frequency and detail: Evidence from a field experiment in customer satisfaction. *Journal of Accounting Research*, *55*(5), 1051–1088.

Chakravarty, S., Harrison, G. W., Haruvy, E. E., & Rutström, E. E. (2011). Are you risk averse over other people's money? *Southern Economic Journal*, *77*(4), 901–913.

Chan, E. W. (2018). Promotion, relative performance information, and the Peter Principle. *The Accounting Review*, *93*(3), 83–103.

Charness, G. (2004). Attribution and reciprocity in an experimental labor market. *Journal of Labor Economics*, *22*(3), 665–688.

Charness, G., Gneezy, U., & Henderson, A. (2018). Experimental methods: Measuring effort in economics experiments. *Journal of Economic Behavior and Organization*, *149*, 74–87.

Chase, W. G., & Simon, H. A. (1973). The mind's eye in chess. In *Visual Information Processing* (pp. 215–281). Elsevier.

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.

Choi, J., Hecht, G., Tafkov, I. D., & Towry, K. L. (2016). Vicarious learning under implicit contracts. *The Accounting Review*, *91*(4), 1087–1108.

Chun, J. S., Brockner, J., & De Cremer, D. (2018). How temporal and social comparisons in performance evaluation affect fairness perceptions. *Organizational Behavior and Human Decision Processes*, *145*, 1–15.

Collewet, M., & Sauermann, J. (2017). Working hours and productivity. *Labour Economics*, *47*, 96–106.

Colquitt, J. A., Conlon, D. E., Wesson, M. J., Porter, C. O., & Ng, K. Y. (2001). Justice at the millennium: A meta-analytic review of 25 years of organizational justice research. *Journal of Applied Psychology*, *86*(3), 425.

Cushman, F., Dreber, A., Wang, Y., & Costa, J. (2009). Accidental outcomes guide punishment in a "trembling hand" game. *PloS one*, *4*(8), e6699.

Datar, S., Kulp, S. C., & Lambert, R. A. (2001). Balancing performance measures. *Journal of Accounting Research*, *39*(1), 75–92.

Davis, M. H. (1980). A multidimensional approach to individual differences in empathy. *Working Paper*.

Deason, S., Hecht, G., Tayler, W. B., & Towry, K. L. (2018). Sooner or later: The effects of timing on managers' discretionary weighting of multiple performance measures. *Working Paper, Available at SSRN 2133092*.

Demeré, B. W., Sedatole, K. L., & Woods, A. (2019). The role of calibration committees in subjective performance evaluation systems. *Management Science*, *65*(4), 1562–1585.

Dennis, S. A., Goodson, B. M., & Pearson, C. A. (2020). Online worker fraud and evolving threats to the integrity of MTurk data: A discussion of virtual private servers and the limitations of IP-based screening procedures. *Behavioral Research in Accounting*, *32*(1), 119–134.

de Oliveira, A. C., Smith, A., & Spraggon, J. (2017). Reward the lucky? An experimental investigation of the impact of agency and luck on bonuses. *Journal of Economic Psychology*, *62*, 87–97.

Dohmen, T., Falk, A., Huffman, D., Sunde, U., Schupp, J., & Wagner, G. G. (2011). Individual risk attitudes: Measurement, determinants, and behavioral consequences. *Journal of the European Economic Association*, *9*(3), 522–550.

Du, F., Erkens, D. H., Young, S. M., & Tang, G. (2018). How adopting new performance measures affects subjective performance evaluations: Evidence from eva adoption by chinese state-owned enterprises. *The Accounting Review*, *93*(1), 161–185.

Dye, R. (2004). Strategy selection and performance measurement choice when profit drivers are uncertain. *Management Science*, *50*(12), 1624–1637.

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

Ellington, J. K., & Wilson, M. A. (2017). The performance appraisal milieu: A multilevel analysis of context effects in performance ratings. *Journal of Business and Psychology*, *32*(1), 87–100.

Erkal, N., Gangadharan, L., & Nikiforakis, N. (2011). Relative earnings and giving in a real-effort experiment. *American Economic Review*, *101*(7), 3330–48.

Ernstberger, J., Link, B., Stich, M., & Vogler, O. (2017). The real effects of mandatory quarterly reporting. *The Accounting Review*, *92*(5), 33–60.

Falk, A., Fehr, E., & Fischbacher, U. (2008). Testing theories of fairness—Intentions matter. *Games and Economic Behavior*, *62*(1), 287–303.

Falk, A., & Fischbacher, U. (2006). A theory of reciprocity. *Games and Economic Behavior*, *54*(2), 293–315.

Feather, N. T. (1999). Judgments of deservingness: Studies in the psychology of justice and achievement. *Personality and Social Psychology Review*, *3*(2), 86–107.

Fehr, E., Goette, L., & Zehnder, C. (2009). A behavioral account of the labor market: The role of fairness concerns. *Annual Review of Economics*, *1*(1), 355–384.

Fehr, E., & Schmidt, K. M. (2001). Theories of fairness and reciprocity-evidence and economic applications. *Working Paper, Available at SSRN 264344*.

Feltham, G. A., & Xie, J. (1994). Performance measure congruity and diversity in multi-task principal/agent relations. *The Accounting Review*, *69*(3), 429–453.

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

Fisher, J. G., Maines, L. A., Peffer, S. A., & Sprinkle, G. B. (2005). An experimental investigation of employer discretion in employee performance evaluation and compensation. *The Accounting Review*, *80*(2), 563–583.

Fisher, J. G., Peffer, S. A., Sprinkle, G. B., & Williamson, M. G. (2015). Performance target levels and effort: Reciprocity across single-and repeated-interaction settings. *Journal of Management Accounting Research*, *27*(2), 145–164.

Foss, N., & Stea, D. (2014). Putting a realistic theory of mind into agency theory: Implications for reward design and management in principal-agent relations. *European Management Review*, *11*(1), 101–116.

Fritz, C., Ellis, A. M., Demsky, C. A., Lin, B. C., & Guros, F. (2013). Embracing work breaks. *Organizational Dynamics*, *42*(4), 274–280.

Gächter, S., Huang, L., & Sefton, M. (2016). Combining "real effort" with induced effort costs: The ball-catching task. *Experimental Economics*, *19*(4), 687–712.

Garicano, L. (2000). Hierarchies and the organization of knowledge in production. *Journal of Political Economy*, *108*(5), 874–904.

Garvey, G. T., & Milbourn, T. T. (2006). Asymmetric benchmarking in compensation: Executives are rewarded for good luck but not penalized for bad. *Journal of Financial Economics*, *82*(1), 197–225.

Ghita, R. S. (2021a). Do supervisors reward observable luck? *Working Paper*.

Ghita, R. S. (2021b). Reporting frequency and learning from experience. *Working Paper*.

Gibbons, R., & Henderson, R. (2012). Relational contracts and organizational capabilities. *Organization Science*, *23*(5), 1350–1364.

Gibbs, M., Merchant, K. A., Van der Stede, W. A., & Vargus, M. E. (2004). Determinants and effects of subjectivity in incentives. *The Accounting Review*, *79*(2), 409–436.

Gibson, F. P. (2000). Feedback delays: How can decision makers learn not to buy a new car every time the garage is empty? *Organizational Behavior and Human Decision Processes*, *83*(1), 141–166.

Gilbert, P., Price, J., & Allan, S. (1995). Social comparison, social attractiveness and evolution: How might they be related? *New Ideas in Psychology*, *13*(2), 149–165.

Giraud, F., Langevin, P., & Mendoza, C. (2008). Justice as a rationale for the controllability principle: A study of managers' opinions. *Management Accounting Research*, *19*(1), 32–44.

Goffin, R. D., & Olson, J. M. (2011). Is it all relative? Comparative judgments and the possible improvement of self-ratings and ratings of others. *Perspectives on Psychological Science*, *6*(1), 48–60.

Golman, R., & Bhatia, S. (2012). Performance evaluation inflation and compression. *Accounting, Organizations and Society*, *37*(8), 534–543.

Gong, N., Boh, W. F., Wu, A., & Kuo, T. (2019). Leniency bias in subjective performance evaluation: Contextual uncertainty and prior employee performance. *Emerging Markets Finance and Trade*, 1–15.

Guo, L., Libby, T., Liu, X., & Tian, Y. (2020). Vertical pay dispersion, peer observability, and misreporting in a participative budgeting setting. *Contemporary Accounting Research*, *37*(1), 575–602.

Gupta, A. K., Smith, K. G., & Shalley, C. E. (2013). The interplay between exploration and exploitation. *Academy of Management Journal*, *49*(4), 693–706.

Gurdal, M. Y., Miller, J. B., & Rustichini, A. (2013). Why blame? *Journal of Political Economy*, *121*(6), 1205–1247.

Hannan, R. L. (2005). The combined effect of wages and firm profit on employee effort. *The Accounting Review*, *80*(1), 167–188.

Hannan, R. L., McPhee, G. P., Newman, A. H., & Tafkov, I. D. (2012). The effect of relative performance information on performance and effort allocation in a multi-task environment. *The Accounting Review*, *88*(2), 553–575.

Hannan, R. L., Rankin, F. W., & Towry, K. L. (2010). Flattening the organization: The effect of organizational reporting structure on budgeting effectiveness. *Review of Accounting Studies*, *15*(3), 503–536.

He, Z. L., & Wong, P. K. (2004). Exploration vs. exploitation: An empirical test of the ambidexterity hypothesis. *Organization Science*, *15*(4), 481–494.

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.

Hecht, G., Maas, V. S., & Van Rinsum, M. (2021). The effects of group incentives and transparency on managers' strategic promotion behavior. *Working Paper, Available at SSRN 3116797*.

Hellmann, T. F., & Thiele, V. (2011). Incentives and innovation:A multi-tasking approach. *American Economic Journal: Microecomics*, *3*(1), 78–128.

Hofmann, C., & Indjejikian, R. J. (2018). Authority and accountability in hierarchies. *Foundations and Trends in Accounting*, *12*(4), 298–403.

Holmstrom, B. (1982). Moral hazard in teams. *The Bell Journal of Economics*, *13*(2), 324–340.

Holt, C. A., & Laury, S. K. (2002). Risk aversion and incentive effects. *American Economic Review*, *92*(5), 1644–1655.

Holzmeister, F. (2017). oTree: Ready-made apps for risk preference elicitation methods. *Journal of Behavioral and Experimental Finance*, *16*, 33–38.

Judge, T. A., & Ferris, G. R. (1993). Social context of performance evaluation decisions. *Academy of Management Journal*, *36*(1), 80–105.

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.

Kajüter, P., Klassmann, F., & Nienhaus, M. (2019). The effect of mandatory quarterly reporting on firm value. *The Accounting Review*, *94*(3), 251–277.

Kaplan, S. E., Petersen, M. J., & Samuels, J. A. (2012). An examination of the effect of positive and negative performance on the relative weighting of strategically and non-strategically linked balanced scorecard measures. *Behavioral Research in Accounting*, *24*(2), 133–151.

Kaplan, S. E., Petersen, M. J., & Samuels, J. A. (2018). Further evidence on the negativity bias in performance evaluation: When does the evaluator's perspective matter? *Journal of Management Accounting Research*, *30*(1), 169–184.

Kelly, K. O., Webb, R. A., & Vance, T. (2015). The interactive effects of ex post goal adjustment and goal difficulty on performance. *Journal of Management Accounting Research*, *27*(1), 1–25.

Konow, J. (2003). Which is the fairest one of all? A positive analysis of justice theories. *Journal of Economic Literature*, *41*(4), 1188–1239.

Krishnan, R., Luft, J. L., & Shields, M. D. (2005). Effects of accounting-method choices on subjective performance-measure weighting decisions: Experimental evidence on precision and error covariance. *The Accounting Review*, *80*(4), 1163–1192.

Lahuis, D. M., & Avis, J. M. (2007). Using multilevel random coefficient modeling to investigate rater effects in performance ratings. *Organizational Research Methods*, *10*(1), 97–107.

Lam, C. F., DeRue, D. S., Karam, E. P., & Hollenbeck, J. R. (2011). The impact of feedback frequency on learning and task performance: Challenging the "more is better" assumption. *Organizational Behavior and Human Decision Processes*, *116*(2), 217–228.

Lee, F., Edmondson, A., Thomke, S., & Worline, M. (2004). The mixed effects of inconsistency on experimentation in organizations. *Organization Science*, *15*(3), 310–326.

Lee, S., & Meyer-Doyle, P. (2017). How performance incentives shape individual exploration and exploitation: Evidence from microdata. *Organization Science*, *28*(1), 19–38.

Libby, R., & Luft, J. (1993). Determinants of judgment performance in accounting settings: Ability, knowledge, motivation, and environment. *Accounting, Organizations and Society*, *18*(5), 425–450.

Libby, T., Salterio, S. E., & Webb, A. (2004). The balanced scorecard: The effects of assurance and process accountability on managerial judgment. *The Accounting Review*, *79*(4), 1075–1094.

Lipe, M. G., & Salterio, S. E. (2000). The balanced scorecard: Judgmental effects of common and unique performance measures. *The Accounting Review*, *75*(3), 283–298.

Luft, J., Shields, M. D., & Thomas, T. F. (2016). Additional information in accounting reports: Effects on management decisions and subjective performance evaluations under causal ambiguity. *Contemporary Accounting Research*, *33*(2), 526–550.

Lurie, N. H., & Swaminathan, J. M. (2009). Is timely information always better? The effect of feedback frequency on decision making. *Organizational Behavior and Human Decision Processes*, *108*(2), 315–329.

Maas, V. S., Van Rinsum, M., & Towry, K. L. (2012). In search of informed discretion: An experimental investigation of fairness and trust reciprocity. *The Accounting Review*, *87*(2), 617–644.

Maas, V. S., & Verdoorn, N. (2017). The effects of performance report layout on managers' subjective evaluation judgments. *Accounting and Business Research*, *47*(7), 731–751.

Manso, G. (2011). Motivating innovation. *The Journal of Finance*, *66*(5), 1823–1860.

March, J. G. (1991). Exploration and exploitation in organizational learning. *Organizational Science*, *2*(1), 71–87.

McKinnon, R. (2013). Getting luck properly under control. *Metaphilosophy*, *44*(4), 496–511.

Mehlhorn, K., Newell, B. R., Todd, P. M., Lee, M. D., Morgan, K., Braithwaite, V. A., ... Gonzalez, C. (2015). Unpacking the exploration–exploitation tradeoff: A synthesis of human and animal literatures. *Decision*, *2*(3), 191.

Merchant, K. A. (1987). How and why firms disregard the controllability principle. In *Accounting and management: Field study perspectives* (pp. 316–338). Harvard Business School Press Boston.

Merchant, K. A., & Otley, D. T. (2006). A review of the literature on control and accountability. *Handbooks of Management Accounting Research*, *2*, 785–802.

Mertins, L., Salbador, D., & Long, J. H. (2013). The outcome effect–a review and implications for future research. *Journal of Accounting Literature*, *31*(1), 2–30.

Miller, D. T., & Ross, M. (1975). Self-serving biases in the attribution of causality: Fact or fiction? *Psychological Bulletin*, *82*(2), 213.

Moers, F. (2005). Discretion and bias in performance evaluation: The impact of diversity and subjectivity. *Accounting, Organizations and Society*, *30*(1), 67–80.

O'Neill, T. A., Goffin, R. D., & Gellatly, I. R. (2012). The use of random coefficient modeling for understanding and predicting job performance ratings: An application with field data. *Organizational Research Methods*, *15*(3), 436–462.

Palan, S., & Schitter, C. (2018). Prolific.ac—A subject pool for online experiments. *Journal of Behavioral and Experimental Finance*, *17*, 22–27.

Pan, X., & Xiao, E. (2016). It's not just the thought that counts: An experimental study on the hidden cost of giving. *Journal of Public Economics*, *138*, 22–31.

Pencavel, J. (2015). The productivity of working hours. *The Economic Journal*, *125*(589), 2052–2076.

Prendergast, C., & Topel, R. (1993). Discretion and bias in performance evaluation. *European Economic Review*, *37*(2-3), 355–365.

Rahmandad, H. (2008). Effect of delays on complexity of organizational learning. *Management Science*, *54*(7), 1297–1312.

Rajan, M. V., & Reichelstein, S. (2006). Subjective performance indicators and discretionary bonus pools. *Journal of Accounting Research*, *44*(3), 585–618.

Rand, D. G., Fudenberg, D., & Dreber, A. (2015). It's the thought that counts: The role of intentions in noisy repeated games. *Journal of Economic Behavior & Organization*, *116*, 481–499.

Rubin, J., & Sheremeta, R. (2015). Principal–agent settings with random shocks. *Management Science*, *62*(4), 985–999.

Ruxton, G. D. (2006). The unequal variance t-test is an underused alternative to Student's t-test and the Mann–Whitney U test. *Behavioral Ecology*, *17*(4), 688–690.

Sebald, A., & Walzl, M. (2014). Subjective performance evaluations and reciprocity in principal-agent relations. *Scandinavian Journal of Economics*, *116*(2), 570–590.

Sezer, O., Zhang, T., Gino, F., & Bazerman, M. H. (2016). Overcoming the outcome bias: Making intentions matter. *Organizational Behavior and Human Decision Processes*, *137*, 13–26.

Shanteau, J. (1988). Psychological characteristics and strategies of expert decision makers. *Acta Psychologica*, *68*(1-3), 203–215.

Simons, R. (2010). *Accountability and control as catalysts for strategic exploration and exploitation: Field study results* (Tech. Rep.). Harvard Business School Boston, MA, USA.

Skowronski, J. J., & Carlston, D. E. (1989). Negativity and extremity biases in impression formation: A review of explanations. *Psychological Bulletin*, *105*(1), 131.

Sloof, R., & Van Praag, C. M. (2010). The effect of noise in a performance measure on work motivation: A real effort laboratory experiment. *Labour Economics*, *17*(5), 751–765.

Stock, J. H., & Watson, M. W. (2015). *Introduction to Econometrics.* Pearson.

Taylor, A., & Greve, H. R. (2006). Superman or the fantastic four? Knowledge combination and experience in innovative teams. *Academy of Management Journal*, *49*(4), 723–740.

Tversky, A., & Kahneman, D. (1981). The framing of decisions and the psychology of choice. *Science*, *211*(4481), 453–458.

Uhlmann, E. L., Pizarro, D. A., & Diermeier, D. (2015). A person-centered approach to moral judgment. *Perspectives on Psychological Science*, *10*(1), 72–81.

Voußem, L., Kramer, S., & Schäffer, U. (2016). Fairness perceptions of annual bonus payments: The effects of subjective performance measures and the achievement of bonus targets. *Management Accounting Research*, *30*, 32–46.

Vroom, V. H. (1964). *Work and Motivation.* Wiley.

Wagenhofer, A. (2014). Trading off costs and benefits of frequent financial reporting. *Journal of Accounting Research*, *52*(2), 389–401.

Weber, R., Camerer, C., Rottenstreich, Y., & Knez, M. (2001). The illusion of leadership: Misattribution of cause in coordination games. *Organization Science*, *12*(5), 582–598.

Wooldridge, J. M. (2007). Inverse probability weighted estimation for general missing data problems. *Journal of Econometrics*, *141*(2), 1281–1301.

Wooldridge, J. M. (2016). *Introductory Econometrics: A Modern Approach*. Nelson Education.

Yang, Y., Onderstal, S., & Schram, A. (2016). Inequity aversion revisited. *Journal of Economic Psychology*, *54*, 1–16.

Yzerbyt, V. Y., Muller, D., & Judd, C. M. (2004). Adjusting researchers' approach to adjustment: On the use of covariates when testing interactions. *Journal of Experimental Social Psychology*, *40*(3), 424–431.

Zell, E., & Alicke, M. D. (2009). Self-evaluative effects of temporal and social comparison. *Journal of Experimental Social Psychology*, *45*(1), 223–227.

Zhang, J., Hedden, T., & Chia, A. (2012). Perspective-taking and depth of theory-of-mind reasoning in sequential-move games. *Cognitive Science*, *36*(3), 560–573.

Zubanov, N. (2012). Risk aversion and effort in an incentive pay scheme with multiplicative noise: Theory and experimental evidence. *ERIM Report Series Reference No. ERS-2012-005-STR*.

Zureich, J. (2020). The effect of using narrow versus broad performance measures on employee decision making. *Working Paper*.

Accounting Information and Discretionary Evaluations

English Summary

Technological advancements now allow companies to report additional information to supervisors. In this dissertation, I examine how middle-level supervisors integrate additional information in their discretionary evaluation decisions, and how employees react to these evaluations. The findings presented in chapter 2 suggest that supervisors are more likely to reward unsuccessful exploration when they receive more frequent performance reports from their employees. In turn, employees do not appear to anticipate this and do not explore more when reporting frequency increases. The findings presented in chapter 3 suggest that supervisors with a wider span of control increase the rewards allocated to top performers and decrease the rewards allocated to the weakest performers. In turn, employees do not anticipate this and do not exert more effort when the span of control widens. The results of chapters 2 and 3 suggest that only changing how supervisors evaluate employees is not enough to change employee behavior because employees do not always anticipate how supervisors will evaluate them. The results of chapter 4 suggest that supervisors reward observable luck because they find it fair to do so. In turn, employees decrease their contribution when supervisors reward observable luck but only after employees learn how supervisors evaluate them through repeated interactions. These results suggest fairness concerns can diminish one of the intended benefits of allowing discretionary evaluations. Specifically, fairness concerns can prevent supervisors from using all available non-contractible information to decrease the weight of luck in employees' compensation.

I examine the research questions outlined in this dissertation using data from case-based and interactive experiments. I collect data from both student-participants and online-participants. All data in this dissertation is available upon request. This dissertation contains three stand-alone studies, where one is co-authored with my supervisor Victor Maas.

Accounting Information and Discretionary Evaluations

Dutch Summary

Technologische vooruitgang stelt bedrijven in staat om additionele informatie aan supervisors te rapporteren. In deze dissertatie onderzoek ik hoe mid-level supervisors deze additionele informatie integreren in hun discretionaire evaluaties en hoe werknemers op deze evaluaties reageren. De bevindingen gepresenteerd in hoofdstuk 2 suggereren dat supervisors meer geneigd zijn niet succesvolle exploratieve activiteiten te belonen wanneer ze meer frequent prestatie rapporten van hun werknemers ontvangen. Werknemers lijken dit echter niet te anticiperen en ondernemen niet vaker exploratieve activiteiten wanneer de frequentie van prestatie rapporten toeneemt. De bevindingen gepresenteerd in hoofdstuk 3 suggereren dat supervisors met een bredere span of control de beloning voor top prestaties verhogen en de beloning van slechte prestaties verlagen. Wederom wordt dit niet door werknemers geanticipeerd en werken ze niet harder onder een supervisor met een bredere span of control. De resultaten van hoofdstukken 2 en 3 suggereren daarom dat enkel het veranderen van hoe supervisors hun werknemers evalueren niet genoeg is om het gedrag van werknemers te veranderen aangezien werknemers niet de wijze waarop ze geëvalueerd worden anticiperen. De resultaten gepresenteerd in hoofdstuk 4 suggereren dat supervisors observeerbare toeval belonen aangezien ze dit rechtvaardig vinden. In reactie hierop verminderen werknemers hun inzet wanneer ze, na herhaaldelijke interactie, achterhalen hoe supervisors evalueren. Deze resultaten suggereren dat zorgen betreft rechtvaardigheid een van de voordelen van discretionaire evaluaties teniet doen, aangezien deze zorgen supervisors ervan weerhouden om alle non-conctractible informatie te gebruiken om de weging van toeval in werknemer compensatie te verlagen.

Ik onderzoek de bovengenoemde vragen door gebruik te maken van data verkregen uit case-based en interactieve experimenten, verkregen aan de hand van zowel student- en onlineparticipanten. Alle data in deze dissertatie is op aanvraag beschikbaar. Deze dissertatie bestaat uit drie op zichzelf staande studies, één waarvan gezamenlijk met mijn supervisor Victor Maas is uitgevoerd.