Forced Rating Systems from Employee and Supervisor Perspectives

Cardinaels, Eddy; Feichter, Christoph

Published in:
Journal of Accounting Research

DOI:
10.1111/1475-679X.12388

Published: 01/01/2021

Document Version
Publisher's PDF, also known as Version of record

Journal of Accounting Research

DOI: 10.1111/1475-679X.12388
Journal of Accounting Research
Vol. 59 No. 5 December 2021
Printed in U.S.A.

CHICAGO BOOTH

Forced Rating Systems from Employee and Supervisor Perspectives

EDDY CARDINAELS *, † AND CHRISTOPH FEICHTER ‡

Received 9 December 2019; accepted 22 May 2021

ABSTRACT

Many firms use forced rating systems in which supervisors must evaluate employees according to a predefined distribution. We develop new theory suggesting that forced ratings are less likely to enhance performance when supervisors assess subjective dimensions of employee performance (e.g., creativity), but can have some harmful side effects. In a laboratory experiment, employees work on a creative task, and supervisors rate their performance. We do not find any difference in the employees' performance or effort in a creative task setting between forced and free ratings. We do, however, find that forced ratings create higher stress for employees (ex post stress scales and biomarkers). Higher stress in turn mitigates the positive effect of effort on creativity.

*Tilburg University; †KU Leuven; ‡WU Vienna University of Economics and Business

Accepted by Luzi Hail. We thank the editor, the anonymous associate editor, and the anonymous reviewer for their many helpful suggestions. We further want to thank Jasmijn Bol, Willie Choi, Isabella Grabner, Christoph Hörner, Anne Lillis, Dieter Smeulders, Naomi Soderstrom, Michael Williamson, Jacob Zureich, seminar participants at Maastricht University, Tilburg University, KU Leuven, the Management Accounting Reading Group at the University of Illinois at Urbana-Champaign, University of Melbourne, University of Amsterdam, and WU Vienna, as well as conference participants at the Dutch Accounting Research Conference 2019, ERMAC 2019, and the Midyear Meeting of the MAS 2020, for their helpful comments. Moreover, we thank Nina Kupper and Tom Smeets for offering us valuable advice on research related to stress measurement. The authors thank the University Fund Limburg and the research theme Culture, Ethics, and Leadership from Maastricht University (former institution of Christoph Feichter) for the generous financial support.

[The copyright line for this article was changed on 2 December after original publication]

Furthermore, we find that actual creativity explains less of supervisors' ratings of employees' performance under forced ratings. Instead, factors that are unrelated to actual creativity, such as eloquent writing and strategic gaming behavior, matter more. Results of an additional online experiment confirm that forced ratings work differently in tasks where performance needs to be evaluated subjectively compared to tasks where objective measures are available.

JEL codes: D91, J33, M40, M41, M52, M55

Keywords: forced rating systems; performance evaluation; creativity; stress and subjectivity

1. Introduction

For bonus allocation and evaluation of employees, firms typically use systems in which supervisors assign performance ratings to employees. To appear fair, reduce confrontation cost, or avoid harm to group cohesion, supervisors often tend to be lenient and insufficiently differentiate employees in these ratings (Kampkötter and Sliwka [2011], Moers [2005], Rynes, Gerhart, and Parks [2005], Bol, Kramer, and Maas [2016]). Consequently, the incentive effect of the ratings diminishes (Prendergast [1999], Moers [2005], Bol [2011]). To overcome these problems, many firms implement forced rating systems, in which supervisors must rate employees according to a predefined distribution (Grote [2005]).[1] However, practitioners say forced ratings can be counterproductive when jobs require more subjective assessments by supervisors, such as judgments on creative work, innovation, or knowledge development. In such settings, forced ratings can lead to excessive stress, to frustration and giving up, and may harm innovation (e.g., Guralnik, Rozmarin, and So [2004]).

We use a creative context that requires a subjective evaluation by the supervisor. We examine how a forced rating system (i.e., supervisors must use the entire rating scale), compared to a free rating system (i.e., supervisors are not restricted in how they assign ratings), affects employees' reactions in terms of effort, stress, and performance and supervisors' rating behavior. This research is important because forced ratings are widely used but also controversial. About one-third of the Fortune 500 firms use some form of a forced rating system to evaluate their employees (Alsever [2008], Bates [2003]). Also human capital–intensive firms, such as audit firms, banks, consultancies, law firms, and tech companies (e.g., Google, Microsoft, Yahoo!), use or have used forced rating systems (*Wall Street Journal* [2014]). Yet forced ratings are purported to be stressful and damaging, particularly

[1] Forced rating systems are sometimes referred to as forced rankings, forced distributions, or rank-and-yank systems (Scullen, Bergey, and Aiman-Smith [2005], Stewart, Gruys, and Storm [2010]). We use the term forced rating systems to refer to all of these systems. In a well-known example of such a forced rating system, at General Electric, 20% of the employees had to receive a rating as top performers, 70% as average, and 10% as lowest ranking (and the last group often had to leave the company) (Stewart, Gruys, and Storm [2010]).

when performance is hard to quantify (e.g., Guralnik, Rozmarin, and So [2004], *Wall Street Journal* [2014]). Despite this being an important and controversial topic, research on it remains scarce. An exception is Berger, Harbring, and Sliwka [2013], who show that forced ratings enhance performance by causing an incentive effect. Although they examine a setting with a clear objective outcome measure, far less is known about contexts in which an objective measure is not available.

We develop new theory to argue that forced rating systems unlikely increase employee performance when supervisors must assess subjective dimensions of employees' work. When performance is assessed subjectively, employees may hold different views of their performance than their supervisors. Moreover, the uncertainty about what is required to achieve the best rating or to avoid the worst can increase their frustration and inclination to give up. Thus, it is not clear whether employees would work harder under forced rating systems, compared to free rating systems, in a more subjective context. The uncertainty and concerns about how to achieve a better rating can, however, cause anxiety about the evaluation, increasing employees' stress (Guralnik, Rozmarin, and So [2004], Rock, Davis, and Jones [2014], Rock [2009]). Therefore, we predict that forced ratings cause high levels of psychological stress for employees. This stress, in turn, can blinker people (Burke [1991], Zak and Nadler [2010]), limiting their state of psychological availability (Binyamin and Carmeli [2010], Byron, Khazanchi, and Nazarian [2010]) and potentially leading to "choking under pressure" (Baumeister [1984]). Thus, although people may work hard on creativity, high levels of stress can undermine their efforts (e.g., Amabile [1996], Webb, Williamson, and Zhang [2013]). Consequently, we hypothesize that high levels of stress mitigate the positive effect of effort on creativity.

For the performance ratings, forced ratings can mitigate leniency and compression in the supervisors' performance evaluations compared to free ratings. This can strengthen the link between a supervisor's ratings of employees and their actual performance. However, rating employees relative to each other or identifying someone as poor performer can be very challenging for supervisors without an objective measure (Rock, Davis, and Jones [2014]). Anecdotes suggest that supervisors use various strategies to cope with this difficulty (e.g., Schleicher, Bull, and Green [2009]), which can weaken the link between actual performance and ratings. For example, they may consider other information about employees or irrelevant dimensions of their output to try to make ratings more objective toward the employees. Moreover, they may strategically game the system by switching the ratings of employees each period to ensure that every employee receives a good and bad rating at some point in time. Given the competing arguments, we do not make a directional prediction about the difference between the forced and free rating systems in the extent to which performance ratings reflect actual performance.

We conduct an experiment in which we match three employee participants with one supervisor participant. Employees develop creative solutions

for societal problems (Cardinaels, Dierynck, and Hu [2020]). Supervisors, whose compensation depends on the creativity of their employees, rate the employees' creative performance on a scale from 1 to 3, where 1 = good performer, 2 = average performer, and 3 = bad performer. After each of the five independent rounds of play, employees receive their ratings. Although these ratings determine the employee's bonus, we also use a time-saving bonus (Tafkov [2013]) to measure employee's willingness to expend costly effort. We measure the creativity of each idea via an independent assessment committee assessing all ideas after all sessions ended (Amabile [1982], Kachelmeier, Reichert, and Williamson [2008]). We manipulate the performance evaluation system (forced vs. free rating system) between participants. In the forced rating system, supervisors must use the entire rating scale, which means that the ratings of 1, 2, and 3 all need to be assigned. The free rating system does not restrict supervisors, which means they can assign the same rating to multiple employees or use the entire spectrum. Besides the effort measure (time spent), we measure the participants' stress levels by psychological stress scales and a biomarker (i.e., cortisol) to capture stress caused by neurological reactions (Binyamin and Carmeli [2010], Byron, Khazanchi, and Nazarian [2010], Dickerson and Kemeny [2004]). For supervisors, we measure their rating behavior.

In contrast to prior research, we do not find higher performance or effort under a forced rating system than under a free rating system in our creative task setting. However, we do find that the forced rating system causes participants to perceive higher levels of stress. Using the sample of male participants, for whom cortisol levels are more sensitive to stress interventions (Kudielka, Hellhammer, and Wüst [2009], Reschke-Hernández et al. [2017]), the results from the cortisol measurement confirm that, compared to free ratings, forced ratings induce neurological reactions that increase stress. Consistent with our prediction, our results further show that greater perceived stress mitigates the positive effects of effort on creativity.

Turning to supervisors' ratings, we find that forced ratings reduce leniency in the ratings. We, however, also find that the creativity of ideas has less impact on the ratings employees receive under forced ratings, compared to free ratings. In line with our theory, supervisors start to value aspects of performance unrelated to creativity that are easier to justify to the employees. Language analysis of the submitted ideas shows that employees' use of eloquent language influence supervisors' ratings more strongly in the forced rating system, even though this is unrelated to actual creativity. Additional tests further show that supervisors strategically game the forced rating system by more often swapping employees to different performance ranks across rounds. These dysfunctional effects can explain why the relation between creativity and ratings is weaker under forced ratings.

We run an additional online experiment with Prolific participants to gain further confidence that forced ratings work differently in settings where performance is evaluated more subjectively, compared to the objective task settings that have been studied before (Berger, Harbring, and

Sliwka [2013]). We again match three employees with one supervisor, who evaluates employees' performance. In a 2 × 2 design, we manipulate the task environment (subjective task vs. objective task) and the performance evaluation scheme the supervisor must use (forced vs. free rating). For the objective task, employees solve slider bars, in which the number of correct sliders provides an objective measure to the supervisor. In contrast, for the subjective task, employees again work on a creative task (i.e., ideas for societal problems), in which an objective measure is not available. The experiment is only a one-round setting, and employees only learn about their rating after the experiment's end. Yet, the results show that forced ratings increase performance in the slider task but not in the creative task. In line with our reasoning, forced ratings increase worries about the evaluation criteria in the creative task but decrease them in the objective task setting. These worries in turn affect the stress that participants experience. Similar to the main experiment, stress decreases the effort–performance relation in the creative task setting. In contrast, in the slider task, stress does not hurt performance, and higher effort directly increases performance. These process results offer corroborating evidence that forced ratings work differently when assessing performance requires subjective judgments.

We contribute to the literature in several ways. First, we contribute to an ongoing debate on forced ratings and show that there are several important problems associated with the use of forced ratings. In contrast to Berger, Harbring, and Sliwka [2013], who studied a setting in which performance can be objectively measured, we do not find a performance-enhancing effect of forced ratings. Instead, when performance must be evaluated subjectively, we find that forced ratings cause significantly higher stress, which weakens the positive connection between effort and creativity. Moreover, while supervisors are rewarded for stimulating creativity, our results suggest that supervisors, under a forced rating system, move away from evaluating creativity per se. Instead, they focus on other aspects of employees' performance that are easier to justify to employees, and they strategically game the system. These detrimental effects may explain why some well-known companies that once used forced ratings have stopped using them (e.g., General Electric, Microsoft, Amazon; *Wall Street Journal* [2014]). Although we test our theory in a creative context, our predictions likely generalize to other aspects of jobs that require a subjective evaluation (e.g., quality, knowledge sharing).

Second, we add to the literature on creativity and control. Although Kachelmeier, Wang, and Williamson [2019] show that incentives can increase the long-term creative performance of individuals, studies typically do not find positive effects on immediate creative performance in more short-term oriented tasks (e.g., Kachelmeier, Reichert, and Williamson [2008], Erat and Gneezy [2016, 2017], Webb, Williamson, and Zhang [2013]; Kachelmeier, Webb, Williamson [2020]). We also consider a short-term task setting and find a positive relation between effort and creative performance when stress is low. However, high stress attenuates this positive

relation. We thus provide direct evidence for the choking-under-pressure argument that these studies often allude to. Moreover, studies often examine the effects of incentives on creative output without having supervisors directly rate employees' creativity (e.g., Kachelmeier, Reichert, and Williamson [2008], Kachelmeier, Wang, and Williamson [2019]). We show the performance evaluation system that supervisors have to use can affect the weights they attach to different aspects of creative performance. Although studies have shown that employees tend to ignore subjective dimensions of performance (Bentley [2019], Choi, Hecht, and Tayler [2012, 2013]), our results suggest that even supervisors may cause such a distortion. These potential distortions can provide a first step in explaining why forced ratings can harm a company's innovation.

Finally, studying the impact of stress caused by incentive systems provides new insights to companies, employees, and society. Studies suggest that about 40 million employees in the European Union experience work-related stress (Parent-Thirion et al. [2007]), and a survey of U.S. and U.K. employees indicates that over a quarter of respondents fear experiencing burnout within the next 12 months (Wrike [2018]). This situation creates tremendous costs for society (e.g., healthcare costs) and for companies (e.g., lack of motivation, absenteeism, and turnover). Our results show that elevated stress undermines the positive effect of effort on creative performance, thereby dampening the employees' output. Gaining insights into how various evaluation systems affect stress using our techniques (e.g., cortisol measurement and stress scales) can help companies to design proper incentives.

2. *Related Literature and Theoretical Predictions*

Performance evaluation systems tend to vary across firms. One key difference in evaluation systems is whether the firm restricts its supervisors to assigning employee ratings from a specific scale such as high, medium, or low performer (Cascio [1991], Dominick [2009]). In a free rating system, supervisors are not restricted and can assign the ratings in any way they find suitable. Even though supervisors can compare the performance of employees, relative to each other, they can provide the same rating to multiple employees in a unit. Research shows that such systems can lead to compression and leniency in the performance ratings (Moers [2005], Bol, Kramer, and Maas [2016], Rynes, Gerhart, and Parks [2005]). To reduce these biases, firms can use a forced rating system, whereby supervisors are required to assign a certain fraction of employees for example to the high-, medium-, or low-performance categories (Dominick [2009]). As supervisors must use the full rating scale, they are forced to evaluate employees relative to each other. The consequences can be severe in practice, where employees receiving the lowest rank are sometimes even dismissed (Lawler [2002], Gupta [2018]). Therefore, such forced rating systems are controversially discussed. Yet many companies use or have used forced ratings,

including knowledge-intensive companies, such as audit firms, consultancies, banks, law firms, and tech companies (*Wall Street Journal* [2014]).

Academic research on forced ratings has been limited. In a simulation study, Scullen, Bergey, and Aiman-Smith [2005] find that the dismissal of lower ranks can lead to improvements in work force potential in the first years of implementation but not in the long run. Schleicher, Bull, and Green [2009] show, in two experiments, that supervisors under a forced rating system perceive assigning ratings as more difficult and have less confidence in their ratings and perceive them as unfair. Finally, Berger, Harbring, and Sliwka [2013] show, in an experiment, that forced rating systems can have positive effects on individuals' performance. They show that forced ratings indeed strengthen the incentive effect of evaluations by reducing compression and leniency biases. These positive effects are observed in a simple task setting where performance can be objectively measured, people of similar ability compete, and sabotage across workers is impossible (Harbring et al. [2007], Berger, Harbring, and Sliwka [2013]).

Yet practitioners argue that forced rating systems can be harmful, particularly for knowledge-intensive companies, where success typically depends on qualities like innovation, citizenship, or creativity, which require subjective judgments for evaluation (Guralnik, Rozmarin, and So [2004], Gupta [2018]). We study the use of forced rating systems, relative to free rating systems, in a creative context in which employees generate creative ideas and evaluators must rate employees according to their creative task performance. We predict that forced ratings may not have the performance-enhancing effects that prior studies documented in objective task settings. Much of the theory that we develop likely also applies to other work environments where employees work on tasks for which output or part of their output is hard to quantify.

2.1 EFFECT OF THE FORCED RATINGS ON EFFORT

Studies show that, in settings with clearly quantifiable measures, inducing higher variation in the ratings between good and bad performers through a forced rating strengthens the incentive effect and hence leads to higher effort, relative to free rating systems (Abeler et al. [2010]; Kampkötter and Sliwka [2011], Berger, Harbring, and Sliwka [2013]). The extent to which this positive effect of forced ratings on the employee's effort also holds in a setting where performance is evaluated subjectively is not clear.

On the one hand, by working harder, employees can improve their performance and the likelihood of receiving a good rating. Consequently, the larger variation induced by the forced rating schemes may increase their effort. On the other hand, the ratings that a supervisor assigns to employees might be debatable because of their subjectivity. In such situations, employees often have more positive views of their own performance than their supervisors do (Alicke et al. [1995]). However, supervisors must assign low ratings in a forced rating system. Given that there is no clear objective signal of why the creative performance of one person is better than that of

another, a relatively low rating can demotivate. Demotivation can also occur because people feel that the higher performer may not deserve the rating. Forced ratings therefore may reduce the motivation to expend effort in future periods. In a free rating system, supervisors can accommodate these concerns by giving, for example, more employees a good rating when more ideas look good or avoiding assigning a good rank to mediocre ideas. Based on this reasoning, forced ratings may lower employee effort. Which effect dominates is unclear. Hence, we formulate a nondirectional hypothesis.

H1: There is a difference in the employees' effort under forced versus free rating systems.

2.2 EFFECT OF THE FORCED RATINGS ON EMPLOYEE STRESS

Practitioners claim that the performance evaluation under forced rating systems can produce adverse effects on stress, which can harm companies (Zak and Nadler [2010]). Psychology theory suggests that two main factors contribute to stress reactions for individuals (Dickerson and Kemney [2004]). First, the situation needs to be one in which employees are concerned about the outcome. That is, the self-identity individuals want to preserve or achieve must be at stake. Second, the situation must include uncontrollability and uncertainty. Individuals cannot fully avoid negative outcomes or succeed with certainty, even though they try to deliver their best efforts. When both of these factors are present, individuals experience stress that can lead to neurological reactions, such as higher levels of cortisol (Dickerson and Kemeny [2004]).

Performance evaluation of individuals satisfies the first factor. As long as individuals work on a task they care about and a supervisor evaluates their work, an individual's self-identity is at stake. We argue that the perceived uncertainty, the second factor, varies with the performance evaluation scheme. Although subjective performance evaluation always entails some uncertainty, which can cause perceptions of injustice, distorted evaluations, and anxiety among employees (Scullen, Bergey, and Aiman-Smith [2005, p. 2], Stewart, Gruys, and Storm [2010], Moon, Scullen, and Latham [2016]), we argue that forced ratings amplify these negative feelings. Under a forced rating system, a supervisor must rank employees. An employee's rating depends on the supervisor's judgment of that person's performance and how the supervisor judges that performance relative to peers. Without objective measures available, this raises concerns about the criteria the supervisor uses for differentiating the employees' performance. With the inherent lack of clear evaluation criteria, this need to rank creates even more uncertainty for employees, which likely causes worry and anxiety about someone's evaluation and the ranking achieved (Hazels and Sasse [2008]). These worries can cause neurological stress reactions in the brain imposing relatively high levels of stress on employees (Binyamin and Carmeli [2010], Byron, Khazanchi, and Nazarian [2010]). Under a free rating scheme, a supervisor still uses a subjective evaluation but the ratings do not directly

depend on how the supervisor evaluates peers' performance. That is, many employees may excel and be rewarded accordingly, largely mitigating concerns about the criteria for differentiation. Consequently, we expect that employees will be less worried and stressed about their evaluation under free ratings. Thus, our second hypothesis predicts that, relative to free ratings, the higher uncertainty of forced ratings contributes to greater stress among employees.[2]

H2: Forced ratings lead to higher employee stress levels than free ratings.

2.3 EFFECT OF EFFORT AND STRESS ON TASK PERFORMANCE

The extent to which the effort and stress induced by forced and free ratings affect task performance is not straightforward. Even though the connection between effort and creativity is weaker than in many other settings, working hard is still one of the drivers of creative task performance (e.g., Amabile [1996], Brüggen, Feichter, and Williamson [2018], Kachelmeier, Wang, and Williamson [2019]). Without sufficient effort and active thinking about creative ideas, high creative performance is difficult to achieve (Brüggen, Feichter, and Williamson [2018], Kachelmeier et al. [2020]). Based on these arguments, one may expect a positive relation between effort and creative task performance.

However, studies argue that the connection between high effort and performance in complex tasks is not obvious (Camerer and Hogarth [1999], Bonner et al. [2000]). Baumeister [1984] shows that high levels of pressure and stress can lead to choking, in which people work hard but their effort does not lead to improved performance. Similarly, Ariely et al. [2009] show that, instead of increasing performance, very high levels of incentives can actually harm performance in certain task settings. In a creative task, such as the one we examine, we argue that excessive stress can also induce choking under pressure, consistent with the argument of Baumeister [1984]. Specifically, we argue that high levels of evaluative stress can affect an individual's state of psychological availability (Burke [1991], Binyamin and Carmeli [2010], Byron, Khazanchi, and Nazarian [2010]), such that effort might not always yield higher creativity.

When people focus too much on achieving high ratings or the criteria for evaluation, they draw valuable resources away from parts of the brain responsible for complex and abstract thinking required for creativity (Rock [2009], Zak and Nadler [2010], Heffernan [2014]) and allocate this attention instead to scoring well on the ratings (Burke [1991], Byron and Carmeli [2010]). Moreover, creativity requires trial and error and risk-taking, which can occasionally lead to failure. People under high levels of

[2] The hypothesis presented is not without tension. In fact, forced ratings could decrease the stress levels of individuals, as there is no chance to renege by giving out only low ratings from the supervisor's perspective.

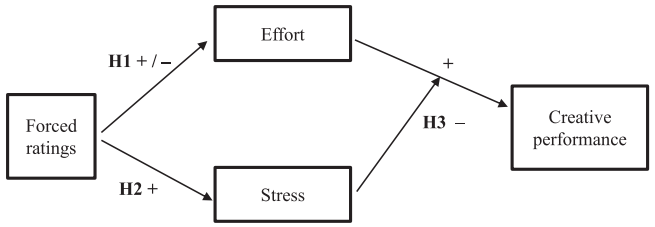

FIG 1.—Theoretical prediction for employee perspective.

stress typically become more risk averse and fall back on habits to avoid a bad rating (Binyamin and Carmeli [2010], Zak and Nadler [2010]), mitigating the extent to which their effort translates into creativity. Based on these insights, we predict that higher levels of stress will mitigate the positive relation between effort and creativity.

H3: High stress levels decrease the positive effect of effort on creativity.

If we summarize our three hypotheses, the effect of the forced rating compared to free rating on overall creative performance is difficult to predict. Although forced ratings may lead to an improvement of effort that we argue to be uncertain in H1, forced ratings at the same time lead to higher levels of evaluative stress as we argue in H2. These higher stress levels can reduce the positive effect that effort has on creative performance as predicted in H3, implying that forced rating systems may have a limited impact on creative performance. Figure 1 summarizes our theoretical predictions from the employee perspective.

2.4 EFFECT OF THE FORCED RATINGS ON THE BEHAVIOR OF RATERS

Although forced ratings pose challenges for supervisors who conduct evaluations (Bates [2003], Rock, Davis, and Jones [2014]), research has largely neglected supervisors' rating behavior. To the extent that free ratings tend to suffer from evaluation biases, forced rating systems can improve the link between actual performance and employee ratings (Berger, Harbring, and Sliwka [2013]). Yet the differentiation of employees can be difficult when a group of employees is small, when realized performance is quite similar, or when objective measures are absent (Lawler [2002], Blume, Baldwin, and Rubin [2009], Schleicher, Bull, and Green [2009], Stewart, Gruys, and Storm [2010], Gupta [2018]). In line with this, Schleicher, Bull, and Green [2009] find that supervisors perceive assigning ratings under a forced system to be very challenging and they do not always perceive these ratings as fair.

In a forced rating system, supervisors must rate employees relative to each other and assign high, medium, and low ratings to their employees. This creates pressure for supervisors to justify their ratings, particularly when assessing performance requires subjective judgment. To overcome

this discomfort, supervisors might resort to alternative strategies, which can distort the link between the actual performance and ratings (Lawler [2002], Bates [2003]). For example, to deal with the justification pressure of their evaluations, supervisors may try to objectify the rating by overemphasizing or including other dimensions of performance that seem easier to justify (e.g., Stewart, Gruys, and Storm [2010], Bol and Smith [2011]). When this happens, a well-described idea that is less novel might be seen as more creative, compared to a more novel one that is not as well-articulated. Or the total sales of a sales agent might be easier to objectively justify than someone's contribution to collegiality and productivity of a group. Consequently, supervisors construct a definition of performance that is easier to justify to their employees at the expense of the "genuine" performance.

Anecdotal evidence further suggests that the pressure to be fair to employees can trigger strategic gaming by supervisors (Schleicher, Bull, and Green [2009]). Supervisors may, for example, strategically swap the highest rating to another individual in each evaluation round to ensure that everyone benefits at some point. The lowest rating may likewise alternate between people when differentiation becomes difficult. These arguments also suggest that the performance of interest might be weighted less under forced ratings than under free ratings.

Given the competing arguments on whether the link between performance and ratings will be improved or distorted under forced ratings, we formulate a nondirectional hypothesis.

H4: The relation between the actual performance and the ratings is different under forced ratings compared to free ratings.

3. Method

For our main laboratory experiment, we recruited 108 volunteers from a participant pool at Maastricht University. The experiment received approval from the ethics committee (i.e., internal review board). Using the z-Tree experimental software (Fischbacher [2007]), we randomly assigned participants to groups of four persons. Within each group, one participant was randomly selected to be the supervisor labeled as player B. The other three players, labeled as Players A1, A2, and A3, played the role of employees, who performed their task individually. Participants stayed in these roles and groups throughout the experiment. On average, participants were 20.6 years old, and 58% were female. The experiment took about 70 minutes. During the experiment, participants could earn experimental points (ECU), which translated to euros once the experiment ended. To keep the expected payoff constant across treatments, we set the conversion rate for each treatment such that the compensation ranged from €5 to €25, with an average payout of €15.6 (Kachelmeier, Reichert, and Williamson [2008]).

3.1 TASKS

3.1.1. Employees. Over the course of five rounds, employees developed creative solutions to societal problems. In each round, we presented employees with a different problem and gave them up to three minutes to develop and describe a creative solution. Participants also had the option to finish early and earn a time bonus (e.g., Tafkov [2013]).[3] In the instructions, we specified that, to be creative, a solution should be *new, innovative, and useful* for solving the problem (Amabile [1996]). Examples of the problem statement are "How to ensure that office workers do more sports" or "How to help smokers quit smoking." Participants could describe as many ideas as they wanted during the three-minute work period. To log and save a solution in the computer program, they needed to press the enter button. Once the three-minute period finished, each participant selected one idea (from the ideas he or she logged during that round) to submit to the supervisor for evaluation.

3.1.2. Supervisors. The supervisors' task was to evaluate the performance of each employee on a scale ranging from 1 to 3, where 1 = good performer, 2 = average performer, and 3 = bad performer. These ratings determined the employees' compensation. To evaluate the ideas, supervisors received the ideas that the three employees submitted for the respective problem after each round. Supervisors could compare the relative creativity of each idea and assign the ratings for each employee on the same page. Together with the original ideas, these ratings were then shown to all three employees, meaning that employees could learn how they performed relative to their peers and which ideas supervisors assessed as more creative in that round.

We selected the task to develop creative solutions to societal problems for several reasons. First, it is an open-ended task without a clear solution, where supervisors need to subjectively assess the performance. Second, employees can use different strategies and focus on different aspects of their ideas, such as the creativity of the idea itself or how it is presented (language) to supervisors. This allows us to disentangle different aspects that supervisors incorporate in their ratings. Third, we can state various problem statements without changing the underlying task characteristics. Finally, the task does not require specific experience or knowledge.

3.2 COMPENSATION

The employees' compensation depends on the performance ratings they receive from their supervisors and a time bonus they could earn. The supervisors' ratings translated into ECU for employees in the following way:

1. Good performer → Employee receives: +300 ECU

[3] The time used to develop the solution and, in particular, whether the employee finished a task before the three minutes was over was never shown to anyone besides the employee.

2. Average performer → Employee receives: +100 ECU
3. Bad performer → Employee receives: –100 ECU

In addition to these points, employees receive 0.5 ECU time bonus for every second they save from the three-minute work period. This means employees face a cost of effort. Working harder increases their chances of receiving the best rating from the supervisor, but it comes at a cost. To determine the total earnings for an employee, we summed up all the points that he or she earned over all five rounds. More ECU leads to higher payoffs.[4]

The supervisors' compensation depends on the creativity of their employees' ideas. This gives supervisors an incentive to rate their employees in such a way as to encourage creativity. Specifically, once the experiment was finished, we showed all the solutions from employees to an independent assessment committee, which assessed the creativity of each solution on a scale from 0 (not very creative) to 100 (very creative). This creativity score directly translates into ECU for supervisors. For every supervisor, we then summed up all the points that the ideas from his or her employees received.

3.3 MANIPULATION

In a 1 × 2 between participants design, we manipulate the performance evaluation scheme supervisors need to use when evaluating their employees (*Forced* vs. *Free* rating). Half of the supervisors have to use a forced rating scheme, which required them to use the entire rating scale. Thus, they must assign each employee a rating of 1 (good performer), 2 (average performer), or 3 (bad performer), without assigning the same rating to another employee. The other half of the supervisors could use the entire range of the scale, but they were not required to do so. They could assign every employee a different rating, but they could also assign the same rating to multiple employees. Given that the roles and groups remained constant throughout the experiment, the supervisor's manipulation also determined the manipulation for the employees, who either worked under a forced or free rating system.

3.4 OUTCOME MEASURES

3.4.1. Effort. To determine the effort, we measure the time used by employees for developing and describing creative ideas in each round (*Time-Spent*) (Bonner et al. [2000]). Remember, participants could stop each round earlier and receive a time bonus for unused time. If they did so, they made less effort to look for solutions that might be better than those they have already conceived.

---

[4] We decided to have a negative number of –100 ECU as the payoff for the lowest rank, because this simulates the fact that forced ratings often come at a cost for the lowest performing employee in a company (e.g., in the extreme case, termination). Remember, however, we set the exchange rate in such a way that the individual with the lowest amount of cumulated ECU still received a payment of €5. Therefore, negative payouts were not possible.

3.4.2. Stress. Employees had to indicate their agreement with the statement "I felt stressed when developing the creative solutions" on a scale from 1 (strongly disagree) to 7 (strongly agree) after every round. This serves as the main variable of stress throughout our analysis (*Stress*). As a second measure, we examine participants' cortisol level, which is a widely used stress indicator in psychology research (e.g., Dickerson and Kemeny [2004] for a meta-study). When individuals experience stress, the human body releases cortisol into the bloodstream. It typically takes between 15 and 25 minutes after a stress intervention for the cortisol to appear in saliva (Dickerson and Kemeny [2004]). Therefore, at three points in time, participants are asked to chew on a swab for 60 seconds to provide samples of their saliva. The first measurement point is before participants start with the main part of the experiment (about 15 minutes after they entered the laboratory), which establishes the individual's base level (*Cortisol1*). The second and third measurements are taken after they finished the main part (with 10 minutes between these two measurements). The average of the second and third measurements therefore picks up the stress caused by the different evaluation schemes in our experiment (*CortEnd*). A higher cortisol level indicates greater stress. The Dresden Lab Service analyzed all saliva samples. Although cortisol gives a measure of stress that is not influenced by perceptions and social desirability, a consistent finding is that the sensitivity of the cortisol measurement via saliva sampling is weaker for females than for males (Dickerson and Kemeny [2004], Reschke-Hernández et al. [2017]). The cortisol response can be influenced by the intake of other hormones (e.g., oral contraceptives) and the menstrual cycles of females. Therefore, we focus on the male population of our sample for our analysis of biomarkers (Kudielka, Hellhammer, and Wüst [2009]).

3.4.3. Creativity. To determine the actual creativity of employees' ideas, which we also use for the supervisor's compensation, we invited an independent assessment committee. Eight different students from the same university completed a 90-minute rating session in the laboratory to assess all the submitted ideas on a scale from 0 (not very creative) to 100 (very creative; e.g., Kachelmeier, Reichert, and Williamson [2008], Brüggen, Feichter, and Williamson [2018]). Participants received €20 for their participation. We presented the problem statement on top of the screen and the solutions they had to assess below. Similar to the main experiment, we instructed these participants that, to be creative, a solution needed to be *new, innovative, and useful* for solving the problem. Once they evaluated all the solutions to one problem statement, they could take a short break and move on to the next page with the next statement. We presented the problem statements in the same order as presented to the employees during the experiment. Moreover, to calibrate the evaluations, we presented the first 30 solutions to each problem in the same order to all eight participants. To keep the amount of ideas manageable, we split the rest of the ideas in such a way that each participant had to assess only half of the remaining ideas.

The Cronbach's alpha for the first 30 solutions to each problem is 0.71, and, for all the other ideas, it is 0.66; both are at acceptable levels (Murphy and Davidshofer [1988]). The creativity score for each idea is equal to the average of the panel's evaluation score (*Creativity*). The mean creativity score assigned by the independent assessment committee was 52.89 (SD = 13.38), ranging from 0.125 to 91 points.[5]

3.4.4. Ratings. The variable of interest for the supervisor perspective is the rating they assign to their employees. For ease of interpretation, we reverse code the ratings such that a zero is the lowest rating that supervisors assign, a one is a medium rating, and a two represents the best rating (*RatingRev*). Thus, a higher number means a better rating.

3.5 PROCEDURE

The experiment consisted of three parts. Before participants started, one of the authors explained the procedure for the cortisol measurement. After participants gave their consent, they moved to the cubicles. During part I, participants responded to demographic questions and personality scales and a 10-item perceived stress scale (*PSS*; Cohen, Kamarck, and Mermelstein [1983], Cohen and Williamson [1988]), which served as a baseline level of perceived stress. In addition, all participants had a fixed three-minute trial round to get familiar with the creative task without any compensation or evaluation. The first saliva sampling then occurred. Next, participants received the paper-based instructions for part II, the main part of the experiment. Before participants learned their role as supervisor or employee, they answered some questions to demonstrate understanding of the instructions (including the manipulation check). Once they answered all questions correctly, they could proceed with the five rounds of the experiment. Each round consisted of a three-minute work period for employees followed by supervisors rating the employees, employees learning about their ratings, and employees and supervisors having to answer questions regarding stress levels and perceived fairness. After the last round, the second cortisol measurement happened. For the third cortisol measurement, participants had to wait for another 10 minutes. During this time, participants completed part III, the postexperimental questionnaire. Finally, participants collected their earnings in the week after the experiment.

4. Results

4.1 EMPLOYEE PERSPECTIVE

Table 1 shows the descriptive statistics for the employee's effort (*TimeSpent*), their *Stress*, and the *Creativity* of their ideas. Given our experimental

[5] In the same week as we ran our main experiment, we ran two additional control treatments described in detail in footnote 17. The independent assessment committee assessed the ideas of all these treatments together (randomly ordered). The above statistics refer to all ideas

TABLE 1
Descriptive Statistics for Employees

		Period					
		1	2	3	4	5	Total
Free rating	*TimeSpent*	125.2	121.8	127.3	123.2	131.9	125.9
		(44.8)	(44)	(44)	(45.4)	(42)	(43.8)
	Stress	3.6	3.8	3.5	3.4	3.4	3.5
		(1.7)	(1.9)	(1.8)	(1.6)	(1.9)	(1.8)
	Creativity	39.7	49	46.9	48.1	49.3	46.6
		(24.9)	(24.6)	(24.7)	(21.6)	(22.1)	(23.7)
	N	42	42	42	42	42	210
Forced rating	*TimeSpent*	125.5	137.9	129.2	135.4	132.6	132.1
		(40.9)	(35.1)	(38.8)	(28.3)	(35.1)	(35.8)
	Stress	4.3	4.1	4.0	3.9	3.7	4.0
		(1.9)	(1.6)	(1.8)	(2.0)	(2.0)	(1.8)
	Creativity	42.8	45	47.9	48.3	49.9	46.8
		(22.7)	(26.6)	(21.6)	(21.4)	(21.6)	(22.8)
	N	39	39	39	39	39	195

This table shows the descriptive statistics (mean and standard deviation) by treatment (forced vs. free rating) of the main dependent variables for the employees' perspective per period. *TimeSpent* = time in seconds that employees spend on the task to create solutions to societal problems (0–180 seconds). *Stress* = response to the question "I felt stressed when developing the creative solutions" on a scale from 1 = strongly disagree to 7 = strongly agree. *Creativity* = score that the independent assessment committee assigned to an idea on a scale from 0 (not very creative) to 100 (very creative). *N* = number of participants.

setup, we collect multiple observations per individual, and participants work in groups that are stable over time.[6] Similar to Berger, Harbring, and Sliwka [2013], we account for this dependency by running random-effect regressions with clustered standard errors at the group level and control for period dummies, unless otherwise stated (Feldman [1988], Angrist and Pischke [2009], Athey and Imbens [2017], Wooldridge [2016]). The random effects account for individual heterogeneity, whereas clustering on group level captures potential differences in group dynamics. The period dummies account for common trends across periods.[7]

---

assessed by the committee. When participants did not submit any idea, we imputed a creativity score of 0.

[6] In our analyses further reported, we treat the data from round 1 as trial and only include the data from rounds two through five for which participants received rating feedback. In round 1, participants might still need to adjust to the experimental setup (i.e., the rating scheme) and try different strategies unrelated to our variables of interest. Our main results are the same if we include round 1 in our analysis.

[7] Our data structure warrants a random-effect estimation with clustered standard errors at a group level. We have 27 group clusters, which should be reasonable, as research suggests between 20 and 50 clusters (Cameron and Miller 2015) for reliable estimates. In untabulated results, we use random-effect regressions and cluster the standard errors on the individual, instead of the group level. While this method captures idiosyncratic individual effects, it does not account for the different group dynamics. The main results comport with those reported in the paper (all p's < 0.10).

TABLE 2
Employees' Perspective

Dependent Variable	(1) *Creativity*	(2) *TimeSpent*	(3) *Stress*	(4) *CortEnd*	(5) *Creativity*
Forced	1.293	4.796	0.942***	0.593*	
	(2.40)	(8.15)	(0.33)	(0.33)	
Stress					2.784
					(1.89)
TimeSpent					0.162***
					(0.04)
TimeSpent_×_Stress					−0.024*
					(0.01)
PSS	0.035	−4.296	0.764***		0.652
	(1.12)	(3.19)	(0.13)		(1.15)
Cortisol1				0.321***	
				(0.11)	
Constant	57.697***	0.000	0.382	1.182**	32.683***
	(5.29)	(0.00)	(0.54)	(0.44)	(8.69)
Observations	283	324	324	35	283
Participants	81	81	81	35	81
R^2	0.021	0.021	0.163	0.319	0.092
Period dummies	Yes	Yes	Yes	No	Yes

This table shows the random-effect regressions for the main variable of interest from the employees' perspective. Standard errors in parentheses are clustered at a group level. Models 1/2/3/5 include the observations from employees in periods 2–5. Model 4 includes only male employees. For model 5, we subtracted the minimum values of *TimeSpent* and *Stress* from these variables, for ease of interpretation. *Forced* = an indicator variable that takes the value of 1 (0) if they worked under the forced rating scheme (free rating scheme). *TimeSpent* = time in seconds that employees spend on the task to create solutions to societal problems (0–180 seconds). *Stress* = the value of the response to the question "I felt stressed when developing the creative solutions." *PSS* = responses to 10-item perceived stress scale from Cohen, Kamarck, and Mermelstein [1983] and Cohen and Williamson [1988]. *Creativity* = score that the independent assessment committee assigned to an idea on a scale from 0 (not very creative) to 100 (very creative). *Cortisol1* = cortisol level of participants before they start the main part of the experiment. *CortEnd* = the mean cortisol level of participants from the second and third cortisol measurement.

***$p < 0.01$, **$p < 0.05$, *$p < 0.1$ indicate significance levels (two-tailed).

We start our analysis with examining how forced ratings, compared to free ratings, affect the employee performance in our experimental task. We run our regression with the *Creativity* of the employees' ideas as the dependent variable and *Forced* as independent variable. As employees in the forced ratings scored significantly lower on Cohen's *PSS* before they started the experiment, we also control for their response on this scale.[8] The results in column 1 of table 2 show that there is no significant difference in the creativity between forced and free ratings ($p = 0.59$). Thus, in contrast to prior studies documenting a positive performance effect of forced ratings in tasks where supervisors have access to an objective measure (e.g., Berger, Harbring, and Sliwka [2013]), we do not find performance enhancing effects in our setting. In the next step, we examine our theoretical model,

[8] Testing for random assignment shows no significant differences between treatments with respect to gender, age, year of study, and risk-taking for supervisors and employees (all p's > 0.10).

which may explain why forced ratings do not enhance the performance in settings requiring a more subjective assessment.

4.1.1. Test of H1: Effect of Forced Ratings on Employee Effort. We predict that forced ratings have a different influence on the employees' effort, compared to free ratings. *TimeSpent* serves as the dependent variable in our regression and the treatment *Forced* as the main independent variable. We also control for *PSS*. Contrary to H1, the results in column 2 show that effort between the *Forced* and free rating systems does not significantly differ (coeff. 4.796, $p = 0.56$). Thus, although studies have shown positive effects of forced ratings on employee effort in settings with objective performance measures, we do not find such a positive effect in our setting.[9]

4.1.2. Test of H2: Effect of Forced Ratings on Employee Stress. H2 predicts that forced ratings lead to higher stress among employees, compared to free ratings. We use the perceived stress as the dependent variable (*Stress*) and *Forced* as the main independent variable and control for *PSS*. The results in column 3 of table 2 show that *PSS* is significantly related to the *Stress* that participants experience during the experiment (coeff. 0.764, $p < 0.01$). More importantly, participants in the forced rating system report higher stress compared to participants in the free rating system (coeff. 0.942, $p < 0.01$), even when we control for the initial differences in the *PSS* in our regression.[10] This provides support for H2.

We also examine the neurological stress reaction using the cortisol level of the male participants. As we only have one observation per person, we run an OLS regression with the cortisol level at the end of the experiment (*CortEnd*) as the dependent variable, the *Forced* rating as the main independent variable, and *Cortisol1* as a control variable to capture the baseline of

---

[9] Analytical studies argue that positive effort effects of forced ratings are stronger when the ability differences within groups are relatively low (e.g., Lazear and Rosen [1981], Hvide [2002]). Berger, Harbring, and Sliwka [2013] formed groups with homogeneous abilities. Based on employees' response to the postexperimental questionnaire item: "In general, I feel that I am good in generating novel ideas" (scale from 1 = strongly disagree to 7 = strongly agree), we median split our sample in groups that show a high (heterogeneous) and low (homogeneous) within-group difference and run separate analyses. Untabulated results show that neither in heterogeneous (coeff. –8.875, $p = 0.13$) nor in homogeneous groups (coeff. 15.789, $p = 0.18$) are there significant differences between forced and free ratings. That said, in line with the argument that employees in heterogeneous groups might get complacent in the forced ratings over time as they figure out that they will win/have no chance to win the tournament, we do find a negative interaction effect of *Forced_×_Period* on effort in the heterogeneous groups (coeff. –10.617, $p < 0.01$), but not in the homogeneous groups (coeff. 2.188, $p = 0.54$). Thus, while we overall do not find support for H1, our analysis shows that forced rating systems work better in homogeneous, compared to heterogeneous groups, by sustaining the employees' effort.

[10] In an untabulated analysis, we run the same regression but also control for the *TimeSpent*. Even though there is a significant relation between the *TimeSpent* and the *Stress* (coeff. 0.004, $p = 0.06$), *Forced* still has a highly significant effect on *Stress* (coeff. 0.922, $p < 0.01$), consistent with H2.

each participant.[11] In line with the results on the perceived stress measure, the results in column 4 show that participants in the forced rating system had a significantly higher cortisol level than participants in the free rating system at the end of the experiment (coeff. 0.593, $p = 0.09$). This provides further evidence that forced rating systems cause greater stress.

4.1.3. Test of H3: Effect of Stress on Relations Between Effort and Creativity. Finally, we predict that high levels of stress mute the positive relation between effort and creativity. We run our regression with the *Creativity* as the dependent variable. The perceived stress measure (*Stress*), the effort measure (*TimeSpent*), and the interaction of stress and effort (*Stress_×_TimeSpent*) are the independent variables. For ease of interpretation, we subtract the minimum values of *Stress* and *TimeSpent* from their values. We also include the *PSS* as control variable. Column 5 shows that *TimeSpent* leads to higher *Creativity* (coeff. 0.162, $p < 0.01$), indicating that creativity increases with effort in our setting. Importantly, however, the negative interaction of *TimeSpent_×_Stress* (coeff. –0.024, $p = 0.06$) shows that the effort–creativity relation is much weaker when participants report higher levels of stress, which is in line with the choking under pressure argument put forward in H3.[12] To gain further insights into this relation, we examine the simple effect of *TimeSpent* at various levels of the stress (minimum, 25th/50th/75th percentile, and maximum). The untabulated results show a significantly positive relation of effort and creativity up to the 50th percentile of stress, but no significant effect at the 75th percentile or maximum anymore. This provides further evidence that effort does not bring additional creative performance when employees are highly stressed.[13]

---

[11] For the subsample of male participants, the mean *CortEnd* is 3.07 nmol/L (SD = 1.22) in the forced ratings treatment and 2.23 nmol/L (SD = 0.66) in the free ratings treatment. The correlation between the perceived stress measure and the cortisol level for this group is 0.161 ($p = 0.05$). As the *Cortisol1* already accounts for baseline differences in the stress among individuals, we do not include the *PSS* in this analysis.

[12] Because of a user error while submitting the ideas, we had to exclude 41 of our 324 observations in this analysis. In some instances, employees described their ideas but did not press the enter button, which meant that their idea was not stored on the computer. We identify these observations by examining when employees worked for more than 50 seconds on the task but did not submit any idea. There is no significant difference of this user error between the free and forced rating employees, and only four employees did not submit any idea more than two times (two in the forced rating treatment and two in the free rating treatment). However, when we include these 41 observations and use a dummy to control for these observations, the results are consistent with stress mitigating the positive effort–creativity relation (*TimeSpent_×_Stress* –0.022, $p = 0.036$).

[13] As we measure both the *TimeSpent* and *Creativity* for our analysis of H3, there might be concerns that we do not document a causal relation but creative individuals simply prefer working longer on the task. While our random-effects model already controls for unobserved heterogeneity in individuals, we can specifically control for individual's creative ability as measured by the PEQ item discussed in footnote 9. Including this variable to our main regression shows that the results remain unchanged. We still find a negative interaction of *TimeSpent_×_Stress* (coeff. –0.024, $p = 0.06$).

In sum, our theoretical model can thus explain why we do not find overall performance effects of forced ratings in our task setting that requires a subjective evaluation. Specifically, we do not find any difference in the effort between forced and free ratings (H1), but we do show that forced ratings lead to higher experienced stress in our setting (H2). High levels of stress in turn mitigate the positive effort–performance relation in our creative task, consistent with the choking under pressure argument (H3). In section 5, we report the results of an additional experiment as corroborating evidence for the different effect that forced ratings in a subjective task setting have on employees, compared with a setting in which an objective measure is available.

4.2 SUPERVISOR PERSPECTIVE

One reason for firms to install forced rating systems is to counteract leniency and compression in the performance ratings. To test whether ratings in a free rating system are more lenient compared to forced ratings, we use the supervisors' ratings as the dependent variable (*RatingRev*) and the *Creativity* of the employees' idea and the *Forced* treatment as the independent variables. Similar to the employee perspective, we run random-effect regressions with clustered standard errors and control for period dummies in all supervisor regressions, unless otherwise stated. We also include a control variable for *NoIdeaSubmitted*, as supervisors had to assign ratings to all three employees, even if they did not submit any idea (but in almost all cases automatically assigned the lowest rating).[14] Results in column 1 of table 3 show that, after we control for the actual *Creativity* score of the ideas, supervisors assign lower ratings in a forced rating system (coeff. –0.254, $p < 0.01$). Hence, a forced system indeed reduces leniency and compression in the performance evaluation.[15] The *Creativity* of ideas themselves has a positive effect on the ratings (coeff. 0.008, $p = 0.02$).

4.2.1. Test of H4: Influence of Creativity on the Ratings. We predict a difference in the relation between actual performance and the ratings supervisors assign between forced and free rating systems. The analysis in column 2 of table 3 shows that creativity indeed has a positive relation with the ratings (coeff. 0.011, $p < 0.01$). However, the significant negative interaction of *Forced_×_Creativity* (coeff. –0.007, $p < 0.01$) shows that supervisors weigh creativity less when determining their ratings in the forced than in the free

[14] We also run the supervisor analyses without this control variable but rather treat them as regular observations. All inferences remain similar. The only *p*-value that would exceed the threshold of 0.10 is the interaction of *Forced_×_FleschKincaid* of column 2 of Table 4, where the two-tailed *p*-value would rise from 0.06 to 0.12.

[15] In an untabulated analysis, we test whether forced ratings decrease compression in the ratings. For each group, we calculate the standard deviation of the ratings per period (*SDRatings*). Running a regression with the *SDRatings* as the dependent variable and *Forced* and the period dummies as independent variables shows that *Forced* significantly increases the variation in the ratings (coeff. 0.262, $p < 0.01$).

TABLE 3
The Effect of Creativity on the Ratings

Dependent Variable	(1) Overall *RatingRev*	(2) Overall *RatingRev*	(3) Forced *RatingRev*	(4) Free *RatingRev*
Creativity	0.008**	0.011***	0.015***	0.008*
	(0.00)	(0.00)	(0.01)	(0.00)
Forced	−0.254***	0.065		
	(0.07)	(0.11)		
Forced_×_Creativity		−0.007***		
		(0.00)		
Period			0.085	−0.057
			(0.07)	(0.04)
Creativity_×_Period			−0.003*	0.001
			(0.00)	(0.00)
NoIdeaReceived	−0.825***	−0.842***	−0.692**	−0.893***
	(0.19)	(0.18)	(0.28)	(0.21)
Constant	0.000	0.000	0.520*	1.084***
	(0.00)	(0.00)	(0.29)	(0.23)
Observations	324	324	156	168
Participants	81	81	39	42
R^2	0.345	0.354	0.220	0.460
Period dummies	Yes	Yes	Yes	Yes

This table shows the random-effect regressions for the relation between the creativity scores of the ideas and the ratings that employees received. Standard errors in parentheses are clustered at a group level. Models 1–4 include periods 2–5. Models 1 and 2 include observations from forced and free rating scheme. Model 3 (4) includes only observations from the forced rating (free rating) scheme. *RatingRev* = reverse coded ratings; 0 is lowest rating and 2 is highest rating. *Creativity* = score that the independent assessment committee assigned to an idea on a scale from 0 (not very creative) to 100 (very creative). *Forced* = an indicator variable that takes the value of 1 (0) if they worked under the forced rating scheme (free rating scheme). *Period* = a continuous variable ranging from 2 to 5. *NoIdeaReceived* = an indicator variable that takes the value of 1 when the employee did not submit any idea.

$^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.1$ indicate significance levels (two-tailed).

rating system. This supports H4. Although forced ratings are supposed to enhance the link between actual creativity and the ratings, this result suggests that the impact of the creativity on the final ratings is lower under forced ratings.

A concern could be that this negative interaction is driven by supervisors who need to differentiate their ratings in the forced systems, even though the creativity of the ideas they need to evaluate is relatively close to each other. Therefore, in each period, we median split our sample based on the standard deviation of the creativity of ideas within a group, and we run the analysis separately for situations in which the creativity of the ideas is close to each other or far apart. Untabulated results show that the interaction of *Forced_×_Creativity* is not significantly different from zero ($p > 0.29$) when the creativity of the ideas is close, but it is significantly negative if the creativity of the ideas is further apart (coeff. –0.007, $p < 0.01$). This suggests that supervisors using the forced ratings fail to adequately rate their employees, particularly in situations in which it would be fairly easy to assign the ratings based on the actual creativity. Moreover, columns 3 and 4 show

that the impact of creativity on the ratings in the forced ratings decreases over time (coeff. –0.003, $p = 0.09$), whereas no such negative time trend occurs in the free rating system ($p = 0.50$). If a mechanical relationship was driving our results, then the negative effect would be stable across periods. This again suggests that supervisors fail to incorporate creativity adequately under forced ratings.[16]

Collectively, these findings support H4; the link between actual performance and the ratings differs with forced ratings. Instead of improving this relation, forced ratings distort it compared to free ratings.[17] We next examine two potential reasons for this distortion. First, supervisors may consider other information that is easier to justify but less relevant to performance; second, they *strategically game* the system by swapping the ratings over time.

4.2.1.1 Influence of Eloquent Language in the Description of Ideas on Ratings. One way in which supervisors may resolve their discomfort when using forced ratings is by focusing on aspects of the performance that are easier to justify to employees. Studies have shown that complex and eloquent language in narratives is often used to impress receivers of a message and obfuscate the content (Merkl-Davies and Brennan [2007], Brennan et al. [2009], Li [2008], Rennekamp [2012], Holoien and Fiske [2013]). Thus, instead of evaluating employees based on the creativity of their ideas, supervisors in the forced ratings might focus more heavily on the language used in these descriptions.

For each idea, we calculate two widely used readability indices. First, we calculate the average word length (the characters per solution divided by the words per solution), where longer words are often perceived to reflect

---

[16] In an alternative specification, we create an ordinal ranking ranging from 1 to 3 based on the creativity score that the ideas within each group and period received from the independent rater panel (*CreativityOrdinal*). We find similar evidence as reported above. If ideas are close together, the interaction of *Forced_×_CreativityOrdinal* is not significant ($p > 0.41$). In contrast, when the creativity of the ideas is further apart and thus when it would be easier to differentiate, we again find a negative interaction of *Forced_×_CreativityOrdinal* (coeff. –0.199, $p = 0.09$), confirming that this connection is again weaker under forced ratings than under free ratings.

[17] To examine whether results would change if participants did not stay in the same group throughout the experiment, we run two additional control treatments with forced and free ratings using 104 different volunteers from the same university. People stay in their roles as employee and supervisor, but they are randomly rematched to new groups after each round. This prevents learning about ability differences or the rating behavior of their supervisors. We rerun our main regressions (with random effects and period dummies, but without group clustering given the rematching) and add an indicator for *Rematching* and an interaction of *Forced_×_Rematching*. Untabulated results show that our main regressions used for testing the hypotheses do not change with adding these two control treatments. Moreover, in none of the regressions the interaction between *Forced_×_Rematching* is significant (all p's> 0.14), and the *Rematching* variable is only significant in the effort regression (coeff. 13.883, $p = 0.05$) suggesting that the rematching groups spend more time on the task (but still no difference between forced and free ratings). Thus, our conclusions are not influenced by whether groups are stable over time or rematched.

TABLE 4

The Influence of Eloquent Language on the Ratings

Dependent Variable	(1) *RatingRev*	(2) *RatingRev*
Creativity	0.010***	0.009***
	(0.00)	(0.00)
Forced	−1.339	−0.353
	(0.86)	(0.27)
Forced_×_Creativity	−0.007***	−0.005**
	(0.00)	(0.00)
WordLength	−0.085	
	(0.07)	
Forced_×_WordLength	0.241*	
	(0.14)	
FleschKincaid		0.005
		(0.01)
Forced_×_FleschKincaid		0.033*
		(0.02)
NoIdeaReceived	−0.887***	−1.002***
	(0.20)	(0.19)
Constant	0.000	0.899***
	(0.00)	(0.24)
Observations	323	323
Participants	81	81
R^2	0.360	0.366
Period dummies	Yes	Yes

This table shows the random-effect regressions for the relation between the creativity scores and the use of eloquent language in the descriptions of the creative solutions and the ratings that employees receive. Standard errors in parentheses are clustered on a group level. Period dummies are included. One observation did not include any text → no readability index. *Creativity* = score that the independent assessment committee assigned to an idea on a scale from 0 (not very creative) to 100 (very creative). *Forced* = an indicator variable that takes the value of 1 (0) if they worked under the forced rating scheme (free rating scheme). *WordLength* = the average word length per idea. *FleschKincaid* = the readability of an idea considering the sentence and word length. *NoIdeaReceived* = an indicator variable that takes the value of 1 when the employee did not submit any idea.

***$p < 0.01$, **$p < 0.05$, *$p < 0.1$ indicate significance levels (two-tailed).

more complex concepts (Lewis and Frank [2016]). Second, we calculate the Flesch–Kincaid grade level index, which gives the U.S. school grade level that is required to understand a text (Kincaid et al. [1975], Benoit, Munger, and Spirling [2019]).[18] For both measures, higher values indicate more eloquent and complex language.

The results in table 4 provide evidence consistent with our reasoning. *Creativity* has a positive relation with the ratings but this effect is weaker under forced ratings (*Forced_×_Creativity* coeff. –0.007, $p < 0.01$ and coeff. –0.005, $p = 0.04$). At the same time, the positive and significant interactions of *Forced_×_WordLength* (coeff. 0.241, $p = 0.10$) and *Forced_×_FleschKincaid* (coeff. 0.033, $p = 0.06$) show that more eloquent language positively

[18] The formula to calculate this index is as follows: [0.35 × (total words/total sentences) + 11.8 × (total syllables/total words) − 15.59]. We used the R package textstat_readability.R to calculate it.

influences the ratings under a forced rating system. Thus, supervisors reduce the weight on the creativity of ideas in favor of other potentially less important dimensions under forced ratings. In fact, an insignificant correlation between *Creativity* and the *WordLength* ($r = 0.036$, $p = 0.48$) and a negative correlation between *Creativity* and *FleschKincaid* ($r = –0.177$, $p < 0.01$) show that the eloquent language does not increase the actual creativity of the ideas.[19]

4.2.1.2 Strategic Gaming by Supervisors. Strategic gaming might be a second explanation for why supervisors in the forced rating system consider the creativity of the ideas less in their ratings, compared to those in the free rating system. To identify this behavior, we count the total number of times a supervisor changes the rating within each group from one period to the other (*SwitchGroup*). The results in column 1 of table 5 show that supervisors change their ratings more frequently in the forced, compared with the free rating treatment (coeff. 1.324, $p = 0.04$). Interestingly, the results in column 2 suggest that this cannot be explained by justified switches from the supervisor. In fact, the negative interaction of *Forced_×_SwitchJustGroup* suggests switches in the group-ratings that would be justified determine the actual switches less in forced, compared with the free ratings (coeff. –0.819, $p = 0.02$).[20]

We also examine how the rating in the last period (*RatingRevLag*) influences the rating in the next period (*RatingRev*), controlling for the actual *Creativity*. Consistent with the notion of strategic gaming, the results in column 3 show that, while last period's rating has no effect on this period's rating in the free rating treatment ($p = 0.15$), there is a significant negative interaction of *Forced_×_RatingRevLag* (coeff. –0.173, $p = 0.08$). Similarly, column 4 shows that, when an employee received the highest rating in the previous period (*WinnerLag*), the likelihood of being the winner in the new period is also significantly lower in the forced rating treatment (coeff. of *Forced_×_WinnerLag* = –0.178, $p = 0.09$). These findings render further support to the claim that supervisors engage in strategic gaming.

---

[19] To gain confidence that supervisors under a forced rating system focus more on objective measures, we ran an MTurk-experiment. Forty supervisors using either forced or free ratings (between participants) had to evaluate 12 employees (matched in groups of three), who performed the Torrance Alternative use task for 2.5 minutes (e.g., develop as many creative ideas for alternative uses of a wine bottle). In this task, supervisors can focus on the quantity of ideas (objective measure) or their creativity. Results confirm that supervisors using forced ratings put a stronger weight on the quantity than under free ratings (coeff. of Quantity_×_Forced 0.031, $p < 0.01$). In contrast, we find no difference with respect to the creativity dimension or the compensation scenario they worked under (variable of fixed compensation).

[20] Responses to postexperimental questionnaire items show that supervisors who swapped the ratings more frequently in the forced ratings (compared to free ratings) feel that their employees consider their rating behavior as fairer (coeff. 0.563, $p = 0.09$) and are aware that their ratings do not fully reflect the creativity of the ideas (coeff. –0.468, $p = 0.11$). This suggests that supervisors strategically switch the ratings.

TABLE 5
Supervisors' Strategic Gaming Behavior

Dependent Variable	(1) *SwitchGroup*	(2) *SwitchGroup*	(3) *RatingRev*	(4) *Winner*
Forced	1.324**	8.147***	−0.067	−0.063
	(0.62)	(2.83)	(0.14)	(0.07)
SwitchJustGroup		0.250		
		(0.20)		
Forced_×_SwitchJustGroup		−0.819**		
		(0.33)		
RatingRevLag			0.075	
			(0.05)	
Forced_×_RatingRevLag			−0.173*	
			(0.10)	
WinnerLag				0.062
				(0.06)
Forced_×_WinnerLag				−0.178*
				(0.11)
Creativity			0.009***	0.004**
			(0.00)	(0.00)
NoIdeaReceived			−0.780***	−0.230*
			(0.19)	(0.12)
Constant	7.214***	5.250***	0.960***	0.247**
	(0.43)	(1.60)	(0.20)	(0.12)
Observations	27	27	324	324
Participants	27	27	81	81
R^2	0.153	0.330	0.352	0.148
Period dummies	No	No	Yes	Yes

This table shows the regressions for the supervisors' strategic switching behavior in the ratings of their employees. Standard errors are in parentheses. Models 1 and 2 are OLS regressions. Models 3 and 4 are random-effects models with clustered standard errors on a group level. *SwitchGroup* = total number of times the supervisor changes the rating within each group from one period to the other period. *RatingRev(Lag)* = reverse coded ratings; 0 is lowest rating and 2 is highest rating in this period (in the previous period). *Winner(Lag)* = indicator variable that takes the value of 1 in case the employee receives the highest rating (received the highest rating in the previous round). *SwitchJustGroup* = total number of times the relative creativity of employees changes from one period to the other period. *Creativity* = score that the independent assessment committee assigned to an idea on a scale from 0 (not very creative) to 100 (very creative).

$^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.1$ indicate significance levels (two-tailed).

5. *Additional Experiment*

A major finding of our main experiment is that forced ratings affect employees differently in our setting, which requires subjective evaluation compared with studies using tasks that can be objectively evaluated (e.g., Berger, Harbring, and Sliwka [2013]). To gain further confidence that forced ratings work differently in these types of settings and gain additional insights into the underlying process, we conduct an online experiment with Prolific participants.[21] In a 2 × 2 between-participants design, we manipulate

[21] Prolific allows the application of pre-screening criteria. We used an approval rate of 95%. Participants needed to be native English speakers, have no literacy problems, and be a minimum of 18 years old. In total, we received 161 responses of which we use 159 in the analyses.

the task environment (objective vs. subjective task) and the rating system (forced vs. free). In the objective task setting, employee participants have three minutes to solve slider bars (in which the number of correctly solved bars is the objective performance indicator), while in the subjective task, employees can use the three minutes to develop a creative idea for a societal problem (i.e., "How to ensure office workers do more sports"). Employees are again matched with two other employees and are made aware that one supervisor (another Prolific participant) will evaluate them later on a scale ranging from 1 to 3, either under a forced or a free rating system, similar to that used in the main experiment. In addition to a starting fee of 2 British pounds (GBP), the lowest rating carried no bonus for employees, the middle rating 5 GBP and the highest rating 10 GBP. To keep the experiment as simple as possible, we run it only for one round, and employees receive information about their rating and their bonus at the moment of their payout (i.e., after the experiment was finished).

If we already find different effects of forced ratings on employee reactions in this simple setup, we can offer more evidence for our reasoning that stress with regard to the evaluation and the consequences on performance differ, depending on the setting. According to our theory, we first expect that forced ratings relative to free ratings cause more uncertainty and worries about the evaluation in the subjective setting. This in turn affects the perceived stress with respect to the evaluation. Finally, we expect that this stress hinders creativity by mitigating the positive effort–performance relation, but we do not expect such a mitigating effect in the objective task setting.

Before going to the analyses on stress measurement, table 6 shows the effect of forced ratings on the performance on both tasks. For the slider task, we measure the number of sliders solved correctly. For assessing the creativity of the ideas, we again recruited an independent assessment committee of eight different Prolific workers, who received 6 GBP.[22] We rank the employees' performance on the respective task, to better compare performance across tasks.[23] The results in column 1 show that forced ratings

---

The two participants we removed failed two of three attention check questions. Participants were on average 31.8 years old and had 12.5 years of work experience. To ensure payment, four supervisors received 5 GBP for evaluating performance of multiple pairs of three employees. Half of the supervisors used a forced rating system, whereas the other half used a free rating system.

[22] We excluded two of the eight raters from the data analysis. One rater assigned a value of 100 to 76 of 81 ideas; the other one had no significant correlation with 5 of the 6 remaining raters and decreases the Cronbach's alpha from 0.736 to 0.619. Our inferences stay the same if we include them in the analysis.

[23] Given the use of an online platform, it is important to control our performance regressions for factors that create variation in online settings. In the slider task, we control for the use of computer mouse that can vary depending on the type of device participants use (i.e. touch screen computers versus desktops with a separate computer mouse). Performance on the slider task is likely to be sensitive to this use. For the creative task, the PEQ item "Internet"

TABLE 6

Performance of Experiment 2

Panel A: Descriptive statistics of the performance

		Rating Scheme		
		Free	Forced	Total
Slider Task	*Number of sliders*	59.7	64.5	62.1
	PerformanceRank	35.9	43.1	39.5
	N	39	39	78
Creative Task	*Creativity*	51.1	49	50
	PerformanceRank	42.9	39.2	41
	N	39	42	81
Total	*PerformanceRank*	39.4	41.1	40.3
	N	78	81	159

Panel B: Regressions for performance

	(1)	(2)	
	Slider Task	Creative Task	Difference
Dependent Variable	*PerformanceRank*	*PerformanceRank*	(1) and (2)
Forced	8.301*	−3.889	chi-square = 2.89,
	(4.92)	(5.41)	*p*-value = 0.*089*
Computer Mouse	13.630***		
	(4.91)		
Internet		−1.803	
		(2.57)	
Constant	36.894***	43.285***	
	(6.00)	(7.14)	
Observations	78	81	
R^2	0.164	0.027	
Gender Control	Yes	Yes	

Panel A shows the descriptive performance results of Experiment 2. Panel B reports the OLS regressions for the performance. Standard errors are in parentheses. Model 1 in panel B reports the results of the slider task. Model 2 in panel B reports the results of the creative task. Difference test of coefficients between columns 1 and 2 is based on seemingly unrelated regression estimation. *PerformanceRank* = the ranked performance on the respective task. The better the performance, the higher the rank. *ComputerMouse* = dummy variable that takes the value of 1 if the participant used a computer mouse, or 0 if not. *Internet* = it is the response to the PEQ item: "To what extent did you use the Internet, books, or asked another individual for help to develop your creative idea?" on the scale from 1 (not at all) to 7 (to a great extent). *Gender Control* = dummy variables for female and other.

$^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.1$ indicate significance levels (two-tailed).

induce higher performance in the slider task, which replicates the findings of Berger, Harbring, and Sliwka [2013]. However, similar to the results of our main experiment, the same forced rating system does not increase performance in the creative task. The difference of the *Forced* coefficients in

captures the extent of use of alternative strategies (Internet, books, help of colleague) to come up with a solution on a scale from 1 (not at all) to 7 (to a great extent). Even though we have no reason to expect an effect on performance, as standard creative answers are not available for the specific societal problem, we still feel it is important to control for this. The use of such alternative strategies is indeed very low (mean = 1.457). Moreover, our statistical inferences remain similar when we exclude either of these controls.

the two regressions is significant (chi-square = 2.89, $p = 0.09$), supporting our claim that the system works differently in a subjective setting.

Table 7 provides the results with respect to the stress and the underlying process. Panel A shows how the rating systems affect antecedents of stress (i.e., stressor) in the different task settings. This stressor measures the worries and uncertainties individuals have with respect to the evaluation. We measure this variable using three postexperimental questionnaire items that capture whether people thought a lot about the criteria, their rating, and their potential for the bonus in the upcoming evaluation while they performed the task (Cronbach's alpha = 0.76). The analysis in model 1 of panel A show that participants under the forced rating worry less about their evaluation than under the free rating (coeff. = –0.718, $p = 0.02$). However, model 2 of panel A shows that this relationship flips for the creative task (coeff. = 0.643, $p = 0.03$), indicating that participants worry more about their evaluation in the forced rating system compared to the free rating system. Consistent with our expectation, the difference of the *Forced* coefficient across the two regressions is significant (chi-square = 10.23, $p < 0.01$).[24] This confirms that the rating systems affect the stressor differently in the subjective versus objective task. Individuals worry more about their evaluation when forced ratings are used in a subjective setting. Importantly, model 3 of panel A confirms that this stressor is an important antecedent for the experienced stress level (coeff. 0.563, $p < 0.01$).[25]

Finally, in panel B of table 7, we examine how stress affects the performance in the different task settings and, in particular, how it affects the effort–performance relation. To measure the effort, we create a variable (i.e., engagement) using three postexperimental questionnaire items, asking participants about their effort, motivation, and engagement in the task (Cronbach's alpha = 0.82). Consistent with our findings of the main experiment (H3), model 2 shows that, while effort has a positive effect on creative performance, the effect is largely mitigated with elevated levels of stress in the creative task (coeff. –5.609, $p < 0.01$).[26] This interaction is not

[24] The lower level of the stressor under forced ratings compared to the free rating in the objective task setting might seem surprising at first. However, it is consistent with the argument that forced ratings can protect employees from the supervisor reneging in the performance evaluation by giving out low ratings to all employees.

[25] To capture the sequential process, we leave out our manipulations in this regression. When we include the variables for our manipulations in the regression with *StressEvaluation* as the dependent variable, they are not significant, as the stressor already absorbs most of the variance of the manipulated factors.

[26] The results show an unpredicted positive main effect of stress on creativity (22.048, $p < 0.01$). We follow up on this observation and calculate the direct effect of stress on creativity at multiple levels of engagement. The results show that already at the 25th percentile of engagement the effect is not significant anymore and the sign is already negative (–0.386, $p = 0.80$). It stays negative and even reaches significance at higher levels of stress (e.g., at the 75th percentile of engagement the effect is –5.994, $p < 0.01$). Untabulated results from experiment 1 show comparable results. This suggests that at minimum effort levels, stress has a positive effect on creativity, which goes away at moderate and high levels of effort.

TABLE 7

Process Model for Experiment 2

Panel A: Influence of treatments on the process variables

	(1)	(2)	(3)	
Dependent Variable	Slider Task *Stressor*	Creative Task *Stressor*	*StressEvaluation*	Difference (1) and (2)
ForcedRating	−0.718**	0.643**		chi-square = 10.23
	(0.31)	(0.30)		*p*-value < 0.01
Stressor			0.563***	
			(0.09)	
Constant	4.780***	3.290***	0.045**	
	(0.33)	(0.30)	(0.44)	
Observations	78	81	159	
R^2	0.070	0.101	0.214	
Gender Control	Yes	Yes	Yes	

Panel B: Influence of process variables on the performance

	(1)	(2)	
Dependent Variable	Slider Task *PerformanceRank*	Creative Task *PerformanceRank*	Difference (1) and (2)
Engagement	5.897*	18.076***	
	(3.47)	(5.46)	
StressEvaluation	3.399	22.048***	
	(6.57)	(6.99)	
Engagement_×_StressEvaluation	−0.793	−5.609***	chi-square = 6.43
	(1.39)	(1.61)	*p*-value 0.*011*
ComputerMouse	14.830***		
	(4.98)		
Internet		−1.981	
		(2.44)	
Constant	14.095	−31.128	
	(15.71)	(23.50)	
Observations	78	81	
R^2	0.171	0.182	
Gender Control	Yes	Yes	

This table shows the regressions for the process model for experiment 2. Standard errors are in parentheses. Model 1 (2) in panel B reports the results of the SliderTask (CreativeTask). Difference test of coefficients between columns 1 and 2 is based on seemingly unrelated regression estimation. Stressor = average of the responses to "While performing the task, I thought a lot about the criteria the Evaluator will use for assigning the ratings," "While performing the task, I thought a lot about the rating I will receive from the Evaluator," "While performing the task, I thought a lot about the potential bonus I can receive," all on a scale from 1 to 7. *StressEvaluation* = response to statement "I feel nervous and 'stressed' about the evaluation the Evaluator will conduct" on a scale from 1 to 7. For ease of interpretation, we subtract the minimum value. *Engagement* = average of the responses to "I put a lot of effort in the task," "I felt motivated to conduct this task," and "I was very much engaged with this task" on a scale from 1 to 7. For ease of interpretation, we subtract the minimum value. *ComputerMouse* = dummy variable that takes the value of 1 if the participant used a computer mouse, or 0 if not. *Internet* = it is the response to the PEQ item: "To what extent did you use the Internet, books, or asked another individual for help to develop your creative idea?" on the scale from 1 (not at all) to 7 (to a great extent). *PerformanceRank* = the ranked performance on the respective task. The better the performance, the higher the rank. *Gender Control* = dummy variables for female and other.

$^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.1$ indicate significance levels (two-tailed).

significant in the objective slider task ($p = 0.57$), where only effort has a positive effect on performance (coeff. 5.897, $p = 0.09$). The difference in the interaction coefficients across these two regressions is significant (chi-square = 6.43, $p = 0.01$). Similar to the main experiment, we examine the effect of engagement on the creative performance at different levels of stress. Consistent with the choking under pressure argument (H3), the untabulated results show that engagement has a significantly positive relation with creativity at the minimum level and 25th percentile of stress, which gets insignificant at the median level and turns negative at the 75th and maximum level of stress.[27]

In sum, this additional experiment shows that forced ratings indeed increase worries about the evaluation (i.e., stressor), which subsequently causes higher stress. This stress has different performance effects in a creative task. Specifically, higher levels of stress reduce the effort–performance relation only in the creative task but not in the slider task, which can explain why we do not observe a beneficial effect on performance from forced rating in the creative task.

6. *Conclusion*

In two experiments, we examine the effects of forced and free rating systems on both employee reaction and supervisor rating behavior. In both experiments, we do not find any difference in the creative task performance between the two rating systems. However, we do find that forced ratings increase employee stress in the creative task setting and that this greater stress reduces the positive effort–creativity relation. In contrast, for a task where objective performance measures are available, we find that a forced rating actually decreases the stress with respect to the evaluation and that a forced rating can lead to performance enhancing effects—similar to findings of prior research (Berger, Harbring, and Sliwka [2013]). This suggests that forced ratings affect employees differently, depending on the task setting. Furthermore, from the supervisor perspective, we show that, even though forced ratings are supposed to increase the link between the actual performance and the ratings employees receive, we find that this relation is actually weaker with forced ratings compared to free ratings. This distortion occurs because supervisors using forced ratings tend to focus on other aspects than the underlying creativity in their evaluations (e.g., eloquent language) and strategically game the system by swapping ratings across individuals and periods. Together these results point to the downsides of forced

[27] Moreover, similar as to our main experiment, we control for participants responses to the PEQ item "In general, I feel that I am good in developing creative ideas" in an untabulated regression. The results are similar to the analysis without the control variable (interaction of *Engagement_×_Stress* coeff. –5.596, $p < 0.01$). This indicates that perceived creative ability unlikely drives the effort–creativity relation.

rating systems in settings where supervisors asses performance more subjectively.

Our findings underline concerns that practitioners often raise with respect to forced ratings. They often argue that such systems are counterproductive, particularly for talent-intensive companies (Guralnik, Rozmarin, and So [2004], Gupta [2018]), where forced ratings may hamper the pursuit of innovation. Our results suggest that forced ratings may not generate performance improvements in jobs where performance is difficult to capture via objective measures (e.g., Campbell [2012]). A caution here is that we obtain our results in a more short-term oriented creativity setting, where incentives may not produce powerful effects (similar as what prior studies have shown). Nevertheless, prior studies using these short-term oriented tasks allude to the choking under pressure argument as explanation why they might not find effects on creative performance, for which we provide direct evidence. Forced rating systems (relative to free ratings) can cause worries about the evaluation, which creates stress (including biological stress reactions). The stress undermines the positive effects of effort on creativity. Also, practitioners warn that forced ratings may create stress. It is important that firms are aware of these potential costs and side effects (e.g., higher stress and supervisors not assigning appropriate ratings). In fact, the greater stress that forced rating systems cause can lead to other long-term side effects, such as higher turnover rates, health problems, and lack of motivation.

Our results also provide avenues for future research. For example, future research can explore how other evaluation systems cause stress or how forced ratings work when other subjective dimensions of performance are evaluated, such as due diligence in an audit world or corporate citizenship or cooperation in knowledge-intensive firms. Research could also explore the effects of such systems in attracting and retaining talent. Some of the adverse effects we document may be mitigated when such systems attract more competitive people. Alternatively, future research can examine the effects of forced ratings in tasks with a longer time horizon where potential stress effects can be mitigated by implementing rest-periods (e.g., Kachelmeier, Wang, and Williamson [2019]). Such research can also examine forced ratings in combination with other control choices (like corporate culture, day off to spend time on innovation) that may induce forced ratings to motivate effort while at the same time alleviate stress that may go along with these systems. Finally, future research can examine the long-term side effects of the distorted ratings that supervisors give in a forced rating system. Firms may be worse off if employees learn that they can sidestep being downgraded by taking less risk or by focusing on less important aspects that matter for winning the tournament.

REFERENCES

Abeler, J.; S. Altmann; S. Kube; and M. Wibral. "Gift Exchange and Workers' Fairness Concerns: When Equality is Unfair." *Journal of the European Economic Association* 8 (2010): 1299–324.

Alicke, M. D.; M. L. Klotz; D. L. Breitenbecher; T. J. Yurak; and D. S. Vredenburg. "Personal Contact, Individuation, and the Better-Than-Average Effect." *Journal of Personality and Social Psychology* 68 (1995): 804–25.

Alsever, J. "What Is Forced Ranking?" 2008. Available at https://www.cbsnews.com/news/what-is-forced-ranking/.

Amabile, T. M. "Social Psychology of Creativity: A Consensual Assessment Technique." *Journal of Personality and Social Psychology* 43 (1982): 997–1013.

Amabile, T. M. *Creativity in Context*. New York: Taylor & Francis Inc., 1996.

Angrist, J. D., and J. S. Pischke. *Mostly Harmless Econometrics: An Empiricist's Companion*. Princeton, NJ: Princeton University Press, 2009.

Ariely, D.; U. Gneezy; G. Loewenstein; and N. Mazar. "Large Stakes and Big Mistakes." *The Review of Economic Studies* 76 (2009): 451–69.

Athey, S., and G. W. Imbens. "The Econometrics of Randomized Experiments," in *Handbook of Economic Field Experiments*. North-Holland: Vol 1 2017: 73–140.

Bates, S. "Forced Ranking." *Hr Magazine* 48 (2003): 62–62.

Baumeister, R. F. "Choking Under Pressure: Self-Consciousness and Paradoxical Effects of Incentives on Skillful Performance." *Journal of Personality and Social Psychology* 46 (1984): 610–20.

Benoit, K.; K. Munger; and A. Spirling. "Measuring and Explaining Political Sophistication through Textual Complexity." *American Journal of Political Science* 63 (2019): 491–508.

Bentley, J. W. "Decreasing Operational Distortion and Surrogation Through Narrative Reporting." *The Accounting Review* 94 (2019): 27–55.

Berger, J.; C. Harbring; and D. Sliwka. "Performance Appraisals and the Impact of Forced Distribution—An Experimental Investigation." *Management Science* 59 (2013): 54–68.

Binyamin, G., and A. Carmeli. "Does Structuring of Human Resource Management Processes Enhance Employee Creativity? The Mediating Role of Psychological Availability." *Human Resource Management* 49 (2010): 999–1024.

Blume, B. D.; T. T. Baldwin; and R. S. Rubin. "Reactions to Different Types of Forced Distribution Performance Evaluation Systems." *Journal of Business and Psychology* 24 (2009): 77–91.

Bol, J. C. "The Determinants and Performance Effects of Managers' Performance Evaluation Biases." *The Accounting Review* 86 (2011): 1549–75.

Bol, J. C., and S. D. Smith. "Spillover Effects in Subjective Performance Evaluation: Bias and the Asymmetric Influence of Controllability." *The Accounting Review* 86 (2011): 1213–30.

Bol, J. C.; S. Kramer; and V. S. Maas. "How Control System Design Affects Performance Evaluation Compression: The Role of Information Accuracy and Outcome Transparency." *Accounting Organizations and Society* 51 (2016): 64–73.

Bonner, S. E.; R. Hastie; G. B. Sprinkle; and S. M. Young. "A Review of the Effects of Financial Incentives on Performance in Laboratory Tasks: Implications for Management Accounting." *Journal of Management Accounting Research* 12 (2000): 19–64.

Brennan, N. M.; E. Guillamon-Saorin; and A. Pierce. "Methodological Insights: Impression Management: Developing and Illustrating a Scheme of Analysis for Narrative Disclosures—A Methodological Note." *Accounting, Auditing & Accountability Journal* 22 (2009): 789–832.

Brüggen, A.; C. Feichter; and M. G. Williamson. "The Effect of Input and Output Targets for Routine Tasks on Creative Task Performance." *The Accounting Review* 93 (2018): 29–43.

Burke, J. "Identity Processes and Social Stress." *American Sociological Review* 56 (1991): 836–49.

Byron, K.; S. Khazanchi; and D. Nazarian. "The Relationship Between Stressors and Creativity: A Meta-Analysis Examining Competing Theoretical Models." *Journal of Applied Psychology* 95 (2010): 201–12.

Camerer, C. F., and R. M. Hogarth. "The Effects of Financial Incentives in Experiments: A Review and Capital-Labor-Production Framework." *Journal of Risk and Uncertainty* 19 (1999): 7–42.

Cameron, C. A., and D.L. Miller. "A Practitioner's Guide to Cluster-Robust Inference." *The Journal of Human Resources* 50 (2015): 317–72

Campbell, D. "Employee Selection as a Control System." *Journal of Accounting Research* 50 (2012): 931–66.

Cardinaels, E.; B. Dierynck; and W. Hu. "Rejections, Incentives, and Employee Creativity: When Chocolate Is Better Than Cash." Working Paper on SSRN, 2020. https://ssrn.com/abstract=3022001.

Cascio, W. F. *Costing Human Resources: The Financial Impact of Behavior in Organizations*. Mason, OH: Thomson South-Western, 1991.

Choi, J. W.; G. W. Hecht; and W. B. Tayler. "Lost in Translation: The Effects of Incentive Compensation on Strategy Surrogation." *The Accounting Review* 87 (2012): 1135–63.

Choi, J. W.; G. W. Hecht; and W. B. Tayler. "Strategy Selection, Surrogation, and Strategic Performance Measurement Systems." *Journal of Accounting Research* 51 (2013): 105–33.

Cohen, S.; T. Kamarck; and R. Mermelstein. "A Global Measure of Perceived Stress." *Journal of Health and Social Behavior* 24 (1983): 385–96.

Cohen, S., and G. Williamson. "Perceived Stress in a Probability Sample of the U.S," in *The Social Psychology of Health: Claremont Symposium on Applied Social Psychology*, edited by S. Spacapam and S. Oskamp. CA: Sage, 1988: 31–67.

Dickerson, S. S., and M. E. Kemeny. "Acute Stressors and Cortisol Responses: A Theoretical Integration and Synthesis of Laboratory Research." *Psychological Bulletin* 130 (2004): 355–91.

Dominick, P. G. "Forced Rankings: Pros, Cons, and Practices," in *The Professional Practice Series. Performance Management: Putting Research into Action*, edited by J. W. Smither and M. London. San Francisco, CA: Jossey-Bass/Wiley, 2009: 411–43.

Erat, S., and U. Gneezy. "Incentives for Creativity." *Experimental Economics* 19 (2016): 269–80.

Erat, S., and U. Gneezy "Erratum to: Incentives for Creativity." *Experimental Economics* 20 (2017): 274–75.

Feldman, H. A. "Families of Lines: Random Effects in Linear Regression Analysis." *Journal of Applied Physiology* 64 (1988): 1721–32.

Fischbacher, U. "z-Tree: Zurich Toolbox for Ready-Made Economic Experiments." *Experimental Economics* 10 (2007): 171–8.

Grote, D. *Forced Ranking: Making Performance Management Work*. Boston, MA: Harvard Business Review Press, 2005.

Gupta, G. "Are You Still Using Force Rankings? Please Stop." *Forbes*, May 23, 2018.

Guralnik, O.; E. Rozmarin; and A. So. "Forced Distribution: Is It Right for You?" *Human Resource Development Quarterly* 15 (2004): 339–45.

Harbring, C.; B. Irlenbusch; M. Kräkel; and M. Selten. "Sabotage in Corporate Contests—An Experimental Analysis." *International Journal of the Economics of Business* 14 (2007): 367–92.

Hazels, B., and C. M. Sasse. "Forced Ranking: A Review." *SAM Advanced Management Journal* 73 (2008): 35–39.

Heffernan, M. A. *Bigger Prize: Why Competition Isn't Everything and How We Do Better*. New York, NY, Simon & Schuster, 2014.

Holoien, D. S., and S. T. Fiske. "Downplaying Positive Impressions: Compensation between Warmth and Competence in Impression Management." *Journal of Experimental Social Psychology* 49 (2013): 33–41.

Hvide, H. K. "Tournament Rewards and Risk Taking." *Journal of Labor Economics* 20 (2002): 877–98.

Kachelmeier, S. J.; B. E. Reichert; and M. G. Williamson. "Measuring and Motivating Quantity, Creativity, or Both." *Journal of Accounting Research* 46 (2008): 341–73.

Kachelmeier, S. J.; L. W. Wang; and M. G. Williamson. "Incentivizing the Creative Process: From Initial Quantity to Eventual Creativity." *The Accounting Review* 94 (2019): 249–66.

Kachelmeier, S. J.; R. A. Webb; and M. G. Williamson. "Do Performance-Contingent Incentives Help or Hinder Divergent Thinking?" Working paper (Presented at the 12th New Direction in Management Accounting Research), 2020.

Kampkötter, P., and D. Sliwka. "Differentiation and Performance: An Empirical Investigation on the Incentive Effects of Bonus Plans." IZA Discussion paper, 2011.

Kincaid, J. P.; R. P. Fishburne, JR.; R. L. Rogers; and B. S. Chissom. *Derivation of New Readability Formulas (Automated Readability Index, Fog Count and Flesch Reading Ease Formula) for Navy Enlisted Personnel*. Springfield, VA: National Technical Information Service, 1975.

Kudielka, B. M.; D. H. Hellhammer; and S. Wüst. "Why Do We Respond so Differently? Reviewing Determinants of Human Salivary Cortisol Responses to Challenge." *Psychoneuroendocrinology* 34 (2009): 2–18.

Lawler, E. E. "The Folly of Forced Ranking." *Strategy + Business* 28 (2002): 28–32.

Lazear, E. P., and S. Rosen. "Rank-Order Tournaments as Optimum Labor Contracts." *Journal of Political Economy* 89 (1981): 841–64.

Lewis, M. L., and M. C. Frank. "The Length of Words Reflects Their Conceptual Complexity." *Cognition* 153 (2016): 182–95.

Li, F. "Annual Report Readability, Current Earnings, and Earnings Persistence." *Journal of Accounting and Economics* 45 (2008): 221–47.

Merkl-Davies, D. M., and N. M. Brennan. "Discretionary Disclosure Strategies in Corporate Narratives: Incremental Information or Impression Management?" *Journal of Accounting Literature* 26 (2007): 116–96.

Moers, F. "Discretion and Bias in Performance Evaluation: The Impact of Diversity and Subjectivity." *Accounting, Organizations and Society* 30 (2005): 67–80.

Moon, S. H.; S. E. Scullen; and G. P. Latham. "Precarious Curve Ahead: The Effects of Forced Distribution Rating Systems on Job Performance." *Human Resource Management Review* 26 (2016): 166–79.

Murphy, K. R., and C. O. Davidshofer. "Psychological Testing," in *Principles, and Applications.* Englewood Cliffs, NJ: Prentice-Hall, 18 1988.

Parent-Thirion, A.; E. F. Macías; J. Hurley; and G. Vermeylen. *Fourth European Working Conditions Survey. European Foundation for the Improvement of Living and Working Conditions.* Luxembourg: Office for Official Publications of the European Communities, 2007.

Prendergast, C. "The Provision of Incentives in Firms." *Journal of Economic Literature* 37 (1999): 7–63.

Rennekamp, K. "Processing Fluency and Investors' Reactions to Disclosure Readability." *Journal of Accounting Research* 50 (2012): 1319–54.

Reschke-Hernández, A. E.; K. L. Okerstrom; A. Bowles Edwards; and D. Tranel. "Sex and Stress: Men and Women Show Different Cortisol Responses to Psychological Stress Induced by the Trier Social Stress Test and the Iowa Singing Social Stress Test." *Journal of Neuroscience Research* 95 (2017): 106–14.

Rock, D. "Managing with the Brain in Mind." *Strategy + Business Magazine* 56 (2009): 58–67.

Rock, D.; J. Davis; and B. Jones. "Kill Your Performance Ratings." *Strategy + Business Magazine* 74 (2014): 1–10.

Rynes, S. L.; B. Gerhart; and L. Parks. "Personnel Psychology: Performance Evaluation and Pay for Performance." *Annual Review of Psychology* 56 (2005): 571–600.

Schleicher, D. J.; R. A. Bull; and S. G. Green. "Rater Reactions to Forced Distribution Rating Systems." *Journal of Management* 35 (2009): 899–927.

Scullen, S. E.; P. K. Bergey; and L. Aiman-Smith. "Forced Distribution Rating Systems and the Improvement of Workforce Potential: A Baseline Simulation." *Personnel Psychology* 58 (2005): 1–32.

Stewart, S. M.; M. L. Gruys; and M. Storm. "Forced Distribution Performance Evaluation Systems: Advantages, Disadvantages and Keys to Implementation." *Journal of Management & Organization* 16 (2010): 168–79.

Tafkov, I. D. "Private and Public Relative Performance Information under Different Compensation Contracts." *The Accounting Review* 88 (2013): 327–50.

Wall Street Journal. "It's Official: Forced Ranking Is Dead." 2014. Available at https://www.wsj.com/articles/its-official-forced-ranking-is-dead-1402372957

Webb, R. A.; M.G. Williamson; and Y. M. Zhang. "Productivity-Target Difficulty, Target-Based Pay, and Outside-The-Box Thinking." *The Accounting Review* 88 (2013): 1433–57.

Wooldridge, J. M. *Introductory Econometrics: A Modern Approach.* Mason, OH: Nelson Education, 2016.

Wrike. "The Stress Epidemic: Employees Are Looking for a Way Out," in *Stress and Productivity Report*. 2018. downloaded from http://www.wrike.com on 6.6.2019.

Zak, P. J., and A. Nadler. "Using Brains to Create Trust: A Manager's Toolbox," in *Neuroeconomics and the Firm*, edited by A. A. Staton, M. Day, and I. M. Welpe. Northhampton, MA: Edward Elgar Publishing, 2010: 66–77.