'Good job!' The impact of positive and negative feedback on performance.

Daniel Goller[1,2], Maximilian Späth[3*]

[1] Centre for Research in Economics of Education, University of Bern

[2] Swiss Institute for Empirical Economic Research, University of St.Gallen

[3] Department of Economics, University of Potsdam

This version: January 30, 2023

Abstract

We analyze the causal impact of positive and negative feedback on professional performance. We exploit a unique data source in which quasi-random, naturally occurring variations within subjective ratings serve as positive and negative feedback. The analysis shows that receiving positive feedback has a favorable impact on subsequent performance, while negative feedback does not have an effect. These main results are found in two different environments and for distinct cultural backgrounds, experiences, and gender of the feedback recipients. The findings imply that managers should focus on giving positive motivational feedback.

Keywords: Feedback, Performance, Causal Analysis, Cultural Background

*Preliminary version. Do not quote or circulate without permission of one of the authors. Comments are very welcome. We like to thank *Swiss-ski* (Swiss ski federation), Michel Roth, David Morris, Alexander Mesch, and the *DSV* (German swimming federation) for valuable insights into the sports contests from the perspective of (former) professional athletes. Also, we thank participants of the ESEA Conference, 2022, and the Berlin School of Economics Workshop, 2022, as well as Enzo Brox, Lisa Bruttel, and Sandro Heiniger for their helpful comments and suggestions.

1 Introduction

Providing performance feedback is one of the main tasks of managers and leaders (Morgeson, DeRue, & Karam, 2010). One important aim of feedback is to create a favorable emotional response. At best, positive or negative feedback can motivate employees and increase their productivity. In the worst case, it leaves the employees frustrated and unproductive. Therefore, the question of how feedback impacts subsequent performance is of tremendous importance.

Consequently, numerous studies investigating the impact of feedback on creativity (Harrison & Rouse, 2015; Itzchakov & Latham, 2020; Kim & Kim, 2020), the learning process of individuals and firms (Hattie & Timperley, 2007; Lee, Lee, & Kim, 2021) or motivation (Deci & Casico, 1972; Fong, Patall, Vasquez, & Stautberg, 2019) emerged. In particular for positive and negative feedback on performance or productivity, studies show the full range from favorable to unfavorable effects (Eggers & Suh, 2019; Kluger & DeNisi, 1996; Podsakoff & Farh, 1989; Sleiman, Sigurjonsdottir, Elnes, Gage, & Gravina, 2020; Waldersee & Luthans, 1994, etc).

The two major difficulties when investigating the impact of feedback on performance are (1) observing truthful and trustworthy feedback in real-incentive situations and (2) quantifying feedback and performance. While observational studies typically fail to satisfactorily tackle the second difficulty, experimental studies cannot fulfill the first requirement. We are not aware of any causal study in which both requirements are met together.

To address this common shortcoming, we exploit a unique setting to estimate the causal effect of positive and negative feedback on subsequent performance. For this purpose, we use data from professional sports: diving as the primary data source, and ski jumping for supplementary analyses. In these sports, individuals' performance is evaluated subjectively by a jury of seven (or five) experienced judges according to precise rules. Each judge independently issues one rating for the task performance (hereafter, "judges rating" or "rating"). Discarding the highest and lowest rating(s), the common assessment of the jury is calculated from the average of the three remaining ratings (hereafter, "jury performance assessment").[1]

Following the definition in Kluger and DeNisi (1996), stating that feedback is information about one's task performance provided by an external agent, we consider the deviation of the discarded (highest and lowest) ratings from the jury's performance assessment as feedback on

[1]Receiving the jury performance assessment can already be seen as a *knowledge of results* (Kluger & DeNisi, 1996) intervention. The analysis of this knowledge of results, however, is beyond the scope of this paper.

task performance. The discarded ratings are not relevant to the assessment of task performance, but this additional information about judges' general perceptions of performance provides feedback that can only work through the motivational channel on subsequent performance. Kluger and DeNisi (1996) argue that the feedback sign depends on the relation between the performance rating and a benchmark. In line with this, discarded ratings define quasi-randomly occurring positive (negative) deviations from the jury performance evaluation that serve as positive (negative) feedback. No deviation from the benchmark implies neutral feedback. We describe the evaluation and feedback process in more detail in Section 3.2, Figure 1.

We test several of the propositions from the model of the seminal work by Kluger and DeNisi (1996) within a single framework. In our setup, the feedback is truthful, accurately observable, and from an external source. Feedback can impact subsequent performance only through its motivational impact. Performance is strongly incentivized and can be precisely quantified. The performance is measured in non-artificial tasks that individuals are not only familiar with but that are routine aspects of their work. What is particularly valuable from a management perspective is that we can investigate the impact of feedback in an international context.

Theoretically guided by the feedback intervention model (Kluger & DeNisi, 1996), we investigate the effect of positive and negative feedback on performance. Further, we investigate the internal and external generalizability of the results. To assess internal generalizability, we can use our extensive data to analyze whether situational (or personal) variables and task characteristics moderate the effects of the feedback intervention on performance. The international sample covering female and male individuals from more than 50 nations from 6 continents offer the unique opportunity to analyze feedback effects for different cultural backgrounds and gender within the same framework. To investigate external generalizability, we complement the main findings with a second, independent setting. We investigate these aspects using both classical statistical and causal machine learning methods. This is followed by analyses examining the feedback interventions' long-term, repetition, and spill-over effects.

Our analysis shows a performance-enhancing causal effect of positive feedback. The favorable effect of positive feedback is found for recipients from different cultural backgrounds, experience levels, and gender. We observe favorable effects even when individuals repeatedly receive positive feedback. The impact of positive feedback is stronger when the relevance of the task is high.

In contrast to all this, negative feedback on average does not have an impact on performance. Merely, the subgroup of the more experienced individuals benefits from negative feedback.

Our findings imply that managers can use positive feedback to enhance the performance of their employees. Importantly, positive feedback can be given repeatedly on a regular basis. It has a favorable impact irrespective of several relevant characteristics of the recipient and can be universally applied in an international context. With our main finding we are in line with the studies conducted by Azmat and Iriberri (2010), Bandiera, Larcinese, and Rasul (2015), Choi, Johnson, Moon, and Oah (2018), and Itzchakov and Latham (2020) for positive feedback and the meta-study by Fong et al. (2019) for negative feedback. We complement decades of research that provides guidelines on how to optimally give feedback (Balcazar, Hopkins, & Suarez, 1985; Alvero, Bucklin, & Austin, 2001; Sleiman et al., 2020).

2 Theoretical framing

To provide a theoretical foundation for the later empirical analysis, we begin by describing the concept of feedback. Then, we collect relevant empirical research and form predictions based on propositions stated by Kluger and DeNisi (1996).

2.1 The concept of feedback

Feedback exists in many forms. Kluger and DeNisi (1996) define feedback as "[...] actions taken by (an) external agent (s) to provide information regarding some aspect (s) of one's task performance" (p. 255). Burgers, Eden, van Engelenburg, and Buningh (2015) distinguish between elaborate and simple feedback. Elaborate feedback typically includes a lengthy explanation, which provides a guide for learning. Simple feedback merely gives information, about whether something was done right or wrong. Burgers et al. (2015) further distinguish between descriptive, comparative, and evaluative feedback. Descriptive feedback – sometimes called objective feedback (Johnson, 2013) – merely sums up behavior shown by the agent. Comparative feedback uses the performance of other individuals as a reference. Evaluative feedback provides a judgment of the performance. Villeval (2020) distinguishes between a cognitive and a motivational perspective. The cognitive perspective rests on the assumption that individuals have imperfect knowledge about their skills. Here, feedback serves as a signal used in an information-updating process. The motivational perspective focuses on the impact of feedback on intrinsic motivation.

Individuals might receive feedback from one agent or several agents. Stone and Stone (1984) find that receiving feedback from two sources instead of one source increases self-perceived task competence. Related, there is a strand of literature analyzing multi-source feedback (Bailey & Fletcher, 2002; Smither, London, & Reilley, 2005), also called 360 degree feedback (DeNisi & Kluger, 2000). Finally, feedback can be with direct consequences or inconsequential. Often feedback comes without direct (monetary) consequences. Still, research shows that agents also react to irrelevant information (Abeler, Falk, Goette, & Huffman, 2011; Cason & Mui, 1998).

The focus of our paper lies on the impact of simple and evaluative feedback on subsequent performance. The feedback is subjective in the sense that is created by subjective evaluation based on objective guidelines. Our study focuses on the impact of single feedback embedded in a multi-source evaluative process. The feedback has no further consequences besides that it can motivate or demotivate the recipient. One important distinction is between positive and negative feedback. We define positive feedback, sometimes called promotion-orientated feedback (Carpentier & Mageau, 2013), as the expression that the evaluated performance is above a certain reference point. We define negative feedback, sometimes called change-orientated feedback (Carpentier & Mageau, 2013) or corrective feedback (Waldersee & Luthans, 1994), as the expression that the rated performance is below the reference.

2.2 Review and hypotheses

In their influential model, Kluger and DeNisi (1996) assume that there are no behavioral effects when there is no discrepancy between the rating and the reference. Positive feedback increases effort if the agent has the possibility to set new self-goals. Likewise, negative feedback leads to an increase in effort. Similarly, Villeval (2020) argues that positive and negative feedback fosters motivation. On the other hand, positive feedback can lead to a decrease in efforts, when individuals have no possibility to set new goals (Kluger & DeNisi, 1996). Negative feedback can discourage individuals when it threatens the self-perception of their competence (Fong et al., 2019).

Some empirical studies show a favorable impact of positive feedback. Choi et al. (2018) find a better performance in a computerized task after purely positive feedback than in a baseline treatment. Itzchakov and Latham (2020) report better performance in a brainstorming task after positive than after neutral feedback. Bandiera et al. (2015) report that positive feedback improves the performance of university students and Azmat and Iriberri (2010) that positive relative rank feedback enhances the performance of high school students. Other studies, such as Podsakoff and Farh (1989) reporting no impact of positive feedback on performance in an object-listing task, find no influence of positive feedback. Waldersee and Luthans (1994) even report an adverse impact of positive feedback on the performance of employees of fast food restaurants.

Empirical work on the effect of negative feedback provides an ambiguous picture. Several studies show a favorable impact of negative feedback. As for positive feedback, Choi et al. (2018) find an improved performance after purely negative feedback in comparison to a baseline treatment. Azmat and Iriberri (2010) find a favorable effect of negative relative rank feedback. Itzchakov and Latham (2020) report a positive impact of negative feedback on performance in a brainstorming task. Podsakoff and Farh (1989) report a favorable impact of negative feedback in an object-listing task. Waldersee and Luthans (1994) find a performance-enhancing effect of negative feedback for employees of fast food restaurants. Some research, such as the meta-study by Fong et al. (2019), shows no impact of negative feedback. Other studies show an unfavorable impact. For example, Deci and Casico (1972) observe that a negative feedback group shows lower motivation to conduct a puzzle task than a control group.

A reason for the ambiguity in reaction to negative feedback might be heterogeneity in the way how individuals update their perception after receiving self-relevant information. Some research finds that agents do not fully update their self-perception after negative information, while they update their self-perception after observing a positive signal (Eil & Rao, 2011; Kuzmanovic, Jefferson, & Vogeley, 2015; Möbius, Niederle, Niehaus, & Rosenblat, 2022; Sharot et al., 2012). This would imply to find no reaction to negative feedback. Yet, other studies observe a rational updating of beliefs for positive and negative information (Barron, 2021) or even an overweighting of negative information (Coutts, 2019; Ertac, 2011), leaving this strand of empirical research inconclusive.

We build our hypotheses on the theoretical model by Kluger and DeNisi (1996). We argue that in the domain of professional performance, there is always the possibility to set more ambitious goals. This indicates that positive feedback might have a favorable impact.

Hypothesis 1 - Positive Feedback:

The performance is better after receiving positive feedback than after receiving neutral feedback.

We follow Kluger and DeNisi (1996) and Villeval (2020) by assuming that also negative feedback has a performance-enhancing effect. We argue that in the field of professional performance, individuals have a rather stable self-perception of confidence.

Hypothesis 2 - Negative Feedback:

The performance is better after receiving negative feedback than after receiving neutral feedback.

A vital aspect that most empirical studies usually can barely answer is the question of the generalizability of these hypotheses. Here, it is useful to distinguish between the two superordinate layers of personal and task-specific characteristics by which effects could be moderated (compare Fong et al. (2019), for example).

For task characteristics, our hypotheses more readily generalize when individuals' responses to feedback are inherently similar irrespective of the difficulty and importance of the task. Difficult and easy tasks might be perceived differently (Moore & Healy, 2008), which can lead to different perceptions of feedback (Pulford & Colman, 1997) and varying subsequent performance (Vancouver & Tischner, 2004). Kluger and DeNisi (1996) argue that the reaction to feedback is stronger the fewer cognitive resources are needed to perform the task. Likewise, performance might differ depending on the importance of the task (Goller & Heiniger, 2022). Here, Kluger and DeNisi (1996) argue that the effectiveness of feedback increases the more attention is on the task. Guided by the model predictions of Kluger and DeNisi (1996), we do not expect generalizability across task characteristics. Accordingly, we expect stronger feedback effects on performance for (relatively) easier tasks needing fewer cognitive resources and more important tasks that require more attention.

Within the personal domain, three potential moderators seem highly relevant in modern workplaces: cultural background, gender, and experience of the feedback recipients. The literature acknowledges that despite the high relevance of cultural differences in a globalized world, non-WEIRD (not coming from Western, Educated, Industrialized, Rich, and Democratic countries) individuals are largely underrepresented in behavioral research (Henrich, Heine, & Norenzayan, 2010). For example, authors postulate differences in self-construals (Markus & Kitayama, 1991), in feedback seeking of individuals (Sully De Luque & Sommer, 2000) and in feedback reaction of firms (Rhee, Alexandra, & Powell, 2020) between collectivistic and individualistic cultures.

Bear, Cushenbery, London, and Sherman (2017) postulate and Berlin and Dargnies (2016), respectively, Roberts and Nolen-Hoeksema (1994) observe different feedback reactions for women than for men. Eggers and Suh (2019) find that the reaction of organizations to negative feedback depends on the experience in the business area. Kluger and DeNisi (1996) propose differential effects for individuals' behavioral or psychological traits. More relevant from a managerial perspective is if those potentially moderating traits are associated with directly observable characteristics of individuals in a company's diverse context. We refrain from forming explicit expectations and leave the question of generalizability for different cultural backgrounds, genders, and experience levels exploratory.

3 Setting and data

We collect data on international competitions of two competitive sports. In the two sports, namely, ski jumping and diving, athletes compete individually in multi-round competitions. In each round, the athletes' task execution is evaluated by multiple professional judges.

Besides the similarities, there are several specifics to each of the sports. In diving, athletes acrobatically jump into the water. We use data on individual performances in three different types of competitions: 1m springboard, 3m springboard, and 10m platform. The scoring consists of two elements. First, each jump is rated by seven judges with respect to the proper execution. Each judge can reward up to 10 style points (in increments of 0.5). The two highest and the two lowest judges' ratings are discarded for the jury performance assessment of the jump, for which the remaining three judges' ratings are summed up. Second, the jury performance assessment is multiplied by the difficulty coefficient, which depends on the complexity of the jump and is assigned to the jump according to the official rules.[2] In competitions between women, points are accumulated over five jumps, and in competitions between men, over six jumps. Depending on the contest there are preliminary rounds and/or semi-finals and the final round.

[2]See `https://resources.fina.org/fina/document/2021/01/12/916f78f6-2a42-46d6-bea8-e49130211edf/2017-2021_diving_16032018.pdf` for a current version of the rules (last accessed on 01/23/2023).

In the winter sport of ski jumping, athletes jump on skis after sliding down a ramp. Scoring consists of four components. First, athletes receive points for the length of their jump. Second, there are compensation points for the force and direction of the wind. Third, scoring depends on the length of the ramp (gate points). Fourth, athletes receive up to 20 style points (in increments of 0.5) for the flight and landing of the jump. The (style) ratings are independently rewarded by five judges according to official rules.[3] The worst and the best rating are discarded and the other three are accounted for the athletes' score of the round. In a typical competition, 50 athletes start in the first round, of which the 30 best reach the final round. After the final round, both jumps' total scores are added to determine the winner and the succeeding rankings.

3.1 Data sets

Table 1: Descriptive statistics

	Diving		Ski jumping	
	Mean	Std. dev.	Mean	Std. dev.
Panel A: Treatments				
Positive Feedback (deviation positive)	0.426	(0.286)	0.316	(0.262)
Negative Feedback (deviation negative)	0.477	(0.320)	0.357	(0.290)
Panel B: Outcomes				
Score	68.737	(14.557)	118.647	(16.204)
Performance (rem. 3 judges' ratings)	7.119	(1.189)	17.771	(0.744)
Performance (all 5 / 7 judges' ratings)	7.110	(1.182)	17.765	(0.741)
Panel C: Covariates				
Compatriot judge	0.248		0.457	
Home event	0.099		0.127	
Experience (Age in years)	22.429	(3.789)	26.836	(4.949)
Female	0.450			
Difficulty	3.211	(0.331)		
Distance			122.608	(11.837)
Prev. Distance			123.940	(11.143)
Prev. Difficulty	3.166	(0.317)		
Prev. Performance	7.270	(0.958)	17.854	(0.580)
N		13075		4529

Notes: Mean and standard deviation (in parentheses; for non-binary variables). rem. = remaining. Some variables were only observed in one of the data sets. Full descriptive statistics in Appendix Table 6.

[3]See `https://assets.fis-ski.com/image/upload/v1665482445/fis-prod/assets/ICR_Ski_Jumping_2022_marked-up.pdf` for a current version of the rules (last accessed on 01/23/2023).

The main analysis is conducted using data on official diving competitions from 2013 through 2017. This includes special events such as World Championships and the Summer Olympics. Except for the first jump, each jump constitutes one observation. We exclude observations where the rating points of the current or subsequent jump are at the lower or upper bound.[4] Athletes who stop competing during the contest are excluded, e.g., due to injury.

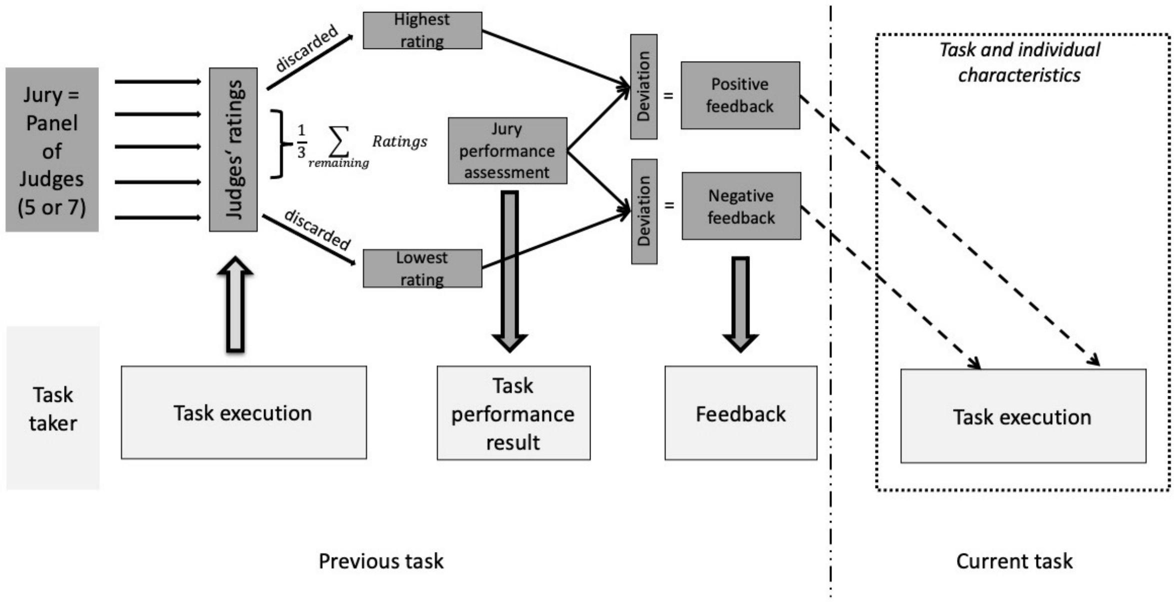
We conduct the analysis based on 13075 observations. The data consists of the jumps performed by 434 athletes from 54 countries in Africa, Asia, Europe, North America, Oceania, and South America. As visible in panel C of Table 1, roughly one-half of the athletes are female and on average 22.4 years old. In 25 percent of the cases, at least one of the judges has the same nationality as the task taker and about 10 percent of observations are at a home event. Difficulty and previous difficulty of the jump are on average around 3.2, and (current and previous) performance are on average around 7.1 to 7.3.

For our analysis on ski jumping, we have 4529 observations on events from the 2010/11 through 2016/17 season (based on a collection conducted by Krumer, Otto, and Pawlowski (2022)). Each observation refers to a second jump. Athletes who fail to qualify for the second round are excluded. In 13 percent of the cases, athletes perform in their respective country of birth. In 45 percent of the cases, one of the judges is of the same nationality as the performing athlete. The average age is about 26.8 years. Jumps are on average about 123 meters and (current and previous) performance are on average around 17.7 (see panels B and C of Table 1).

[4]To put it more concretely: We remove observations that have received an average score of 9.5 or higher (19.5 in ski jumping), as well as those with an average score of less than 5 (14 in ski jumping). Furthermore, we remove observations with individual scores of 3 or lower (14 in ski jumping), as these are most likely to be crashes. All of these choices are robust to changes, and we show the robustness of the results to data pre-processing in the results section.

3.2 Variables

Figure 1: Illustration of the evaluation and feedback process



Notes: For a current task (on the right), feedback is given for the previous task (left). The broken arrows represent our main hypotheses, i.e., the potential influence of feedback on performance in the subsequent task. Task and individual characteristics (dotted square) potentially moderate this effect. In the case of seven judges, the two highest and lowest ratings are discarded, and only the most extreme ratings are used. See Section 5.5 for other specifications used in the robustness checks.

Figure 1 describes the evaluation and feedback process in our setup. For the task execution evaluation, each judge in the jury independently gives a numerical rating for the task execution of the task taker. The largest and smallest of those judges' ratings are discarded and the jury performance assessment is the mean of the remaining (three) judges' ratings. The task performance assessment quantifies the task performance result.

In our study, we focus on the discarded judges' ratings that are not regarded for the jury's performance assessment and can affect subsequent performance only through their motivational impact. Our treatment variables are constructed as deviations of the discarded judges' ratings from the jury performance assessment. More concrete, *Deviation positive* is constructed by subtracting the jury performance assessment (the mean of the ratings in absence of the discarded ratings) from the largest discarded judges' rating. *Deviation negative* is constructed by subtracting the smallest discarded judges' rating from the jury performance assessment.[5] We define

[5]Additionally, we construct and test two alternative specifications. All specifications can be found in the full descriptive statistics in Appendix Table 6. Especially, for diving, there are two (highest/lowest) judges' ratings discarded. The base specification uses the most extreme judges' ratings. Other specifi-

Deviation positive as positive feedback and *Deviation negative* as negative feedback. Panel A in Table 1 provides an overview of the main treatment variables. Both feedback variables, with mean values of 0.426 (0.316) for positive feedback and 0.477 (0.357) for negative feedback, range from 0 (for neutral feedback) to 2.5 (for increasingly positive/negative feedback).

To measure the effect of feedback on subsequent task execution, we use the jury's performance assessment that the task takers receive for their subsequent performance (hereafter, "Performance") as our outcome variable. An alternative variable to measure subsequent performance is the mean of the ratings from all (5 or 7) judges.

4 Empirical strategy

We study how positive and negative feedback affect subsequent performance. To this end, our identification strategy relies on conditional idiosyncratic variations in the differences between the jury performance assessment and the discarded ratings. This positive (negative) deviation is irrelevant to the assessment of the task performance but provides feedback in the form of additional information about the judges' general perception of the performance.

The identification strategy presumes that, once we condition on a few observable characteristics, there are no omitted influences that are correlated with both outcome, i.e., performance in the task, and treatment, i.e., the positive/negative deviation (feedback for the previous task). Our approach formalizes to the following linear baseline model:

$$Y_i = \alpha + \beta_+ A_i^+ + \beta_- A_i^- + \gamma X_i + \epsilon_i,$$

where the outcome, Y_i, is the performance in the (current) task for individual i. The continuous treatments $A_i^{+/-}$ are defined as the positive/negative feedback for the (previous) task, and $\beta_{+/-}$ are the coefficients of interest to investigate our hypotheses 1 and 2. X_i contains (pre-determined) covariates of individual i that we need to control for. ϵ_i is an idiosyncratic error term.

To give credence to the unconfoundedness assumption, we address concerns raised in the literature about potential biases in subjective ratings. First, we consider nationality bias (Heiniger

cation descriptions and results for the robustness of the alternative treatment variable specifications can be found in Section 5.5.

& Mercier, 2021; Krumer et al., 2022; Sandberg, 2018; Zitzewitz, 2006), i.e., a judge from the same country as the task taker rates the compatriot better than other individuals. To account for potentially more positive ratings from judges who are compatriots, we include a) a binary variable indicating whether a judge on the panel is a compatriot of the task taker, and b) an indicator if the individual competes in a home event in X_i.[6] To alleviate remaining concerns about bias based on common nationality, we conduct two further checks. A balancing test in Table 8 shows no balancing issues related to compatriot judges. To ensure that the results are not driven by individuals that are potentially subject to nationality bias, we perform a robustness check in which the affected task takers are removed from the sample.[7]

Second, there is evidence in the literature of an order of action bias (Damisch, Mussweiler, & Plessner, 2006; Ginsburgh & Van Ours, 2003). Subjective ratings are found to be affected by the order of task performance, which threatens our identification when some but not all judges are affected. We account for this by controlling for the order in which individuals perform tasks (starting order). Third, more difficult tasks were found to be rewarded with higher scores–the difficulty bias (Morgan & Rotthoff, 2014). The difficulty of a task in our case is precisely measurable and predetermined. Specifically, in diving, we control for the difficulty of the jump (chosen a priori); in ski jumping, we control for the (previous and current) wind and gate, i.e., the length of the hill–both factors that can influence difficulty and subjective evaluation.

Fourth, there could be reputation bias (Findlay & Ste-Marie, 2004). This bias can lead to better ratings for well-established individuals who typically have a better reputation. To ensure conditional independence, we take into account a) individual and individual-by-season fixed effects and b) current rank in the competition. Fifth, the accuracy of subjective performance ratings is found to vary for different performance qualities (Heiniger & Mercier, 2021). Therefore, we include the individual mean and standard deviation of the jury's performance assessment of the previous task in X_i.

While not testable, we are confident that the conditional independence assumption is satisfied. Still, we offer two types of checks for it. First, in a total of 20 balancing checks in Table 8, only one statistically significant test indicates a solid balancing among observable characteristics. Second, with respect to unobservable characteristics, we provide an indirect approach to sup-

[6]Judges' decisions regarding possible bias in favor of compatriots might be different in front of a supportive crowd (Page & Page, 2010; Goller & Krumer, 2020).

[7]The results for this can be found in Table 11 and hardly differ materially from the main results.

port the conditional independence assumption by implementing a placebo treatment test. We replace the treatment variable with a pseudo-treatment variable recorded in the future. The task performance cannot be influenced by the feedback given in the future of this task. Therefore, if we observe all confounding influences, the placebo treatment effect should be zero. If we reject this placebo null hypothesis this points to some unobserved confounding (or other issues like endogeneity or reverse causality), while not rejecting gives some evidence that the conditional independence assumption is plausible. Table 7 shows that this placebo test cannot reject our assumption of unconfoundedness.

To estimate the main effects of interest, we use linear regression and cluster standard errors on the individual level. In the second step, we apply a method from the causal machine learning literature. For this research, the importance of investigating potential non-linearities in the effect lies in the differently observed treatment intensities, i.e., high or low quantified feedback, for which it is unclear if an estimated constant treatment effect reflects various treatment intensities properly.

With the non-parametric kernel method for continuous treatment effects introduced by Kennedy, Ma, McHugh, and Small (2017) we investigate the effects for different intensities of the treatment. The method builds on two steps. First, a (doubly-robust) pseudo-outcome is constructed as follows:

$$\xi(\pi, \mu) = \frac{Y - \mu(X, A)}{\pi(A|X)} \int \pi(A|x) dP(x) + \int \mu(x, A) dP(x),$$

where the nuisance functions $\pi(A|X)$ and $\mu(X, A)$ are estimated using a random forest estimator (Breiman, 2001). The pseudo-outcome $\xi(\pi, \mu)$ is doubly-robust in the sense that only (at least) one of the two nuisances needs to be consistent, not both, and is free from confounding influences. In the second step, the average potential outcome for given treatment levels is estimated using a non-parametric kernel regression of the pseudo-outcome on the continuous treatment variable: $E(Y^a) = E(\xi(\pi, \mu|A = a))$.

5 Results

5.1 Main results

Our first main finding is that positive feedback is enhancing (subsequent) performance. Panel A in Table 2 shows a statistically significant and positive coefficient for positive feedback. The effect is robust to the inclusion of different sets of covariates. In each specification, the average effects are statistically significant at the 1% level. Panel B replicates this finding for our second data set. As our second main finding, we observe that negative feedback causes an effect close to zero in both panels and all specifications. We do not see any effect of negative feedback on performance.

Table 2: The effect of feedback on performance – sensitivity to different specifications

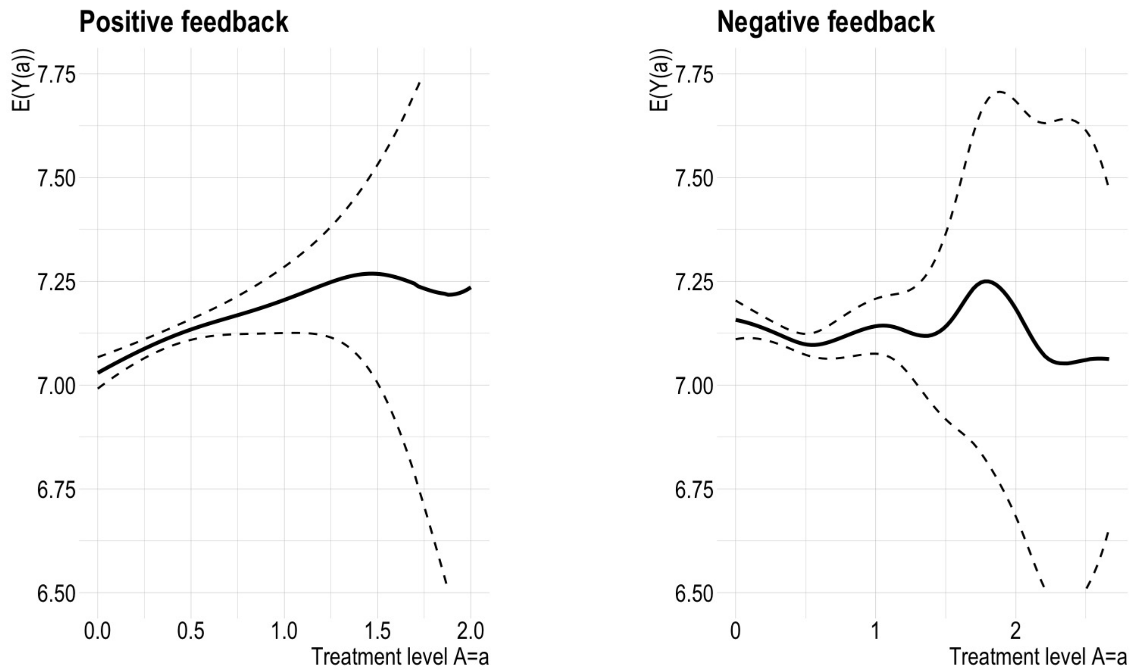
Performance	(1)	(2)	(3)	(4)
Panel A: Diving (N=13075)				
Positive Feedback	0.242***	0.208***	0.115***	0.100***
	(0.036)	(0.034)	(0.032)	(0.035)
Negative Feedback	0.018	0.024	0.001	0.007
	(0.030)	(0.030)	(0.029)	(0.030)
Panel B: Ski jumping (N=4529)				
Positive Feedback	0.201***	0.180***	0.145***	0.107***
	(0.035)	(0.036)	(0.034)	(0.034)
Negative Feedback	-0.063	-0.055	-0.049	-0.026
	(0.043)	(0.041)	(0.037)	(0.041)
Base Covariates	x	x	x	x
All Covariates		x	x	x
Individual Fixed Effect			x	
Individual x Season FE				x

Notes: Linear regression. Full regressions in Tables 9 and 10. All regressions contain previous' jumps jury assessment (*Base Covariates*). *All Covariates* include prev. jumps wind and gate points and distance (ski jumping) or difficulty (diving). Also, points behind, compatriot judge, home event, current ranking, SD of previous performance, and start order. Standard errors are clustered on the individual level. *, **, and *** represents statistical significance at the 10 %, 5 %, and 1 % level, respectively.

The performance-enhancing impact of positive feedback is rather insensitive to the inclusion of more covariates and fixed effects. We start with controlling only for performance in the previous task in column (1). In column (2) we add several control variables as discussed in

Section 4. Columns (3) and (4) add individual fixed effects and individual-by-season fixed effects to the regressions. Detailed result tables can be found in the appendix in Tables 9 and 10, and for the sake of simplicity, all of the following regressions are based on the specification used in column (3).

Figure 2: Non-linear estimation of feedback on performance



Notes: Non-parametric kernel regression for different levels of positive (left) and negative (right). feedback. Expected outcomes (y-axis) and treatment levels (x-axis) are displayed. Kernel bandwidths are 0.300 (left) and 0.214 (right) and are determined in a data-driven approach using a cross-validation method. To obtain treatment effects, one might calculate the difference of the expected outcomes for two treatment levels and divide this by the difference in the treatment levels (treatment intensity). Diving data. The broken lines represent the 90% confidence intervals.

Our results show that, on average, positive feedback is enhancing performance. In the following, we go beyond average effects and investigate the effect of positive and negative feedback for different magnitudes of feedback. Figure 2 provides non-linear estimates of positive and negative feedback showing the expected outcome (performance) against the extent of the feedback, i.e., the level of the treatment. The (treatment) effect of different feedback intensities can be calculated as the difference in expected outcomes for an increase from some treatment level to another.[8] In the graph on the left, the effect of positive feedback is positive throughout all feed-

[8]For two different treatment levels $A = a_1$ and $A = a_0$, the effect can be calculated as $\theta(a_1, a_0) = \frac{E(Y(A=a_1))-E(Y(A=a_0))}{a_1-a_0}$. The treatment intensity in this example is $a_1 - a_0$, while for a complete picture, it needs to be clear that the treatment level from which the treatment intensity is evaluated is a_0 here.

back intensities, i.e., the expected outcome increases almost steadily as the level of treatment increases. With negative feedback, on the right side of Figure 2, the effect varies slightly up and down for different treatment intensities – although the effect does not appear to be different from zero for any treatment intensity, consistent with the average effect of zero reported in Table 2. For both estimations, we find that the linearity assumption in the regression analyses is a good approximation for the non-linear effect curves. Still, especially for the higher treatment intensities the confidence intervals become large and conclusions become imprecise–a fact to which global linear regression models do not give any hint.

Overall, the results provide support for hypothesis 1: The performance is better after receiving positive feedback than after receiving neutral feedback. Contrarily, we do not find support for hypothesis 2, i.e., the performance is not better after receiving negative feedback than after receiving neutral feedback. In the next section, we test if the positive effect of positive feedback and the null effect of negative feedback persists in different sub-populations and is generalizable for diverse personal or situational conditions.

5.2 Sub-population and context heterogeneity

In the feedback-intervention model of Kluger and DeNisi (1996), as well as, for example, in the meta-study of Fong et al. (2019) aspects are collected for which the effects of feedback potentially differ. Personal characteristics, situational aspects, and task characteristics, among other factors, might shape the reaction of individuals to positive and negative feedback. A strength of our unique data set is that it allows us to investigate if we can generalize the results of our analysis.

Panel A of Table 3 exhibits that positive feedback has a favorable impact irrespective of individuals' personal characteristics. We consider three categorizations of the individuals' cultural backgrounds. First, we report that the favorable effect of feedback on performance is present for individuals from WEIRD and non-WEIRD countries. Second, we find a favorable impact of positive feedback irrespective of the relative cultural distance to the U.S.. Third, individuals coming from relatively individualistic and relatively collectivistic countries both react favorably to positive feedback.[9] Other personal characteristics that we investigate are experience and

[9] We classify (non-)WEIRD countries according our own assessment based on Henrich et al. (2010); the respective list can be obtained upon request. For cultural distance to the U.S., we use the metrics provided in Table 1 in the research article by Muthukrishna et al. (2020). For individualistic and

gender. We find a performance-enhancing effect of positive feedback for both the relatively more and less experienced. Similar to Bear et al. (2017), we also explore whether there are gender differences in the reaction to feedback. We find that both sexes react favorably to positive feedback For none of the three different definitions of cultural background, nor gender and experience, do the two-sample WALD tests show statistically significant differences. This leads to the conclusion that the effects of feedback are consistent and generalizable across these three personal characteristics.

Importantly, we find some heterogeneity with respect to the characteristics of the task. Contested situations offer greater incentives to perform (Goller & Heiniger, 2022), with higher task focus and more pressure. Panel B of Table 3 shows large and positive effects for positive feedback in close competitions, but an insignificant effect for situations that are less competitive. This is in line with the argumentation by Kluger and DeNisi (1996) and our expectations. Contrary, we find no support for differential effects for the difficulty of the task. Positive feedback leads to a performance-enhancing impact for easy and hard tasks.

The results of the heterogeneity analysis on the impact of negative feedback are largely in line with the main finding. The second column of Table 3 shows a null effect of negative feedback for most subgroups and all contexts. The only exception is the experience of the individuals, where we find that relatively more experienced individuals improve their performance after receiving negative feedback. A two-sample Wald test (in square brackets) shows that the difference in the reaction between the more and less experienced individuals is statistically significant. The favorable impact of negative ratings for experienced individuals is in line with findings by Eggers and Suh (2019) on the firm level.

collectivistic countries, we use data from the index created by Hofstede (2011).

Table 3: Differential effects

	Positive Feedback	Negative Feedback
Panel A: Individuals' characteristics		
WEIRD[1] (N=4955)	0.086* (0.048)	0.006 (0.049)
Non–WEIRD (N=8120)	0.135*** (0.043)	0.004 (0.037)
	[0.447]	[0.974]
Culturally close to U.S.[2] (N=6223)	0.132*** (0.046)	-0.007 (0.047)
Not culturally close to U.S. (N=6852)	0.101** (0.044)	0.008 (0.037)
	[0.626]	[0.802]
Individualistic country[3] (N=6013)	0.096** (0.047)	0.007 (0.045)
Collectivistic country (N=6872)	0.144*** (0.045)	0.001 (0.040)
	[0.461]	[0.921]
More experienced (age $\geq$ 23y, N=6176)	0.146*** (0.045)	0.076* (0.039)
Less experienced (age $<$ 23y; N=6899)	0.081* (0.047)	-0.062 (0.044)
	[0.318]	[0.019]
Female (N=5885)	0.087* (0.047)	-0.028 (0.042)
Male (N=7190)	0.128*** (0.043)	0.018 (0.039)
	[0.520]	[0.422]
Panel B: Task characteristics		
Tight competition[4] (N=5118)	0.173*** (0.056)	-0.033 (0.052)
Non–tight competition (N=7957)	0.064 (0.039)	0.007 (0.037)
	[0.110]	[0.531]
Easy task[5] (N=7267)	0.154*** (0.043)	-0.027 (0.037)
Hard task (N=5808)	0.086* (0.048)	0.025 (0.044)
	[0.291]	[0.366]

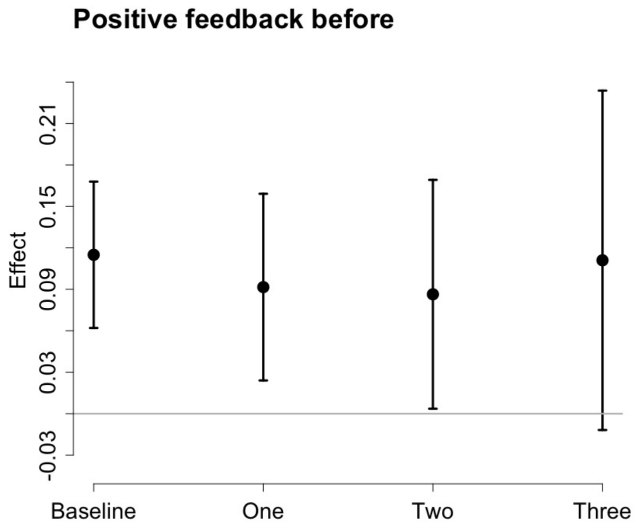
Notes: Linear Regression estimates. Diving data. Control variables as in column (3) in Table 2. Standard errors are clustered on the individual level. *, **, and *** represents statistical significance at the 10 %, 5 %, and 1 % level, respectively. P-value of WALD test for equality in square brackets. [1]Western, Educated, Industrialized, Rich, Democratic. [2]Cultural closeness is divided at the median level of an index taken Muthukrishna et al. (2020). [3]Divided at median level of an individualism index constructed by Hofstede (2011); (some countries missing). [4]Athlete is within ten points to first place in final, and to the cut-off in preliminary rounds. [5]Easy and hard according to the median chosen difficulty of the (assessed) task.

5.3 Repetition and long-term effects

For practitioners, it is crucial to know about the impact of feedback when it is given repeatedly and about its long-term effect. Fortunately, our data allows for analyzing the impact of feedback on performance in a repeated setup.

Figure 3 shows that the favorable impact of positive feedback is non-diminishing with repetition. As a benchmark, *Baseline* shows the average effect of receiving feedback as reported in Table 2, which is not conditional on further previously received feedback. We find that for those who have received positive feedback at least one time before, further positive feedback continues to have a positive impact on their performance. Similarly, we find a positive influence of positive feedback if the individual has received positive feedback at least two or three times before.

Figure 3: A non-diminishing effect of positive feedback



Notes: Linear regression estimates. Diving data. Specifications as in column (3) in Table 2. Standard errors are clustered on the individual level. Effect among those that experienced positive feedback at least one, two, or three times (in the respective round) before. The whiskers mark the 90 % confidence intervals.

Table 4 shows the non-persistence of the effect of positive feedback on performance. For reference, column (1) reports the baseline effect for the performance in the task that is conducted directly after the feedback is received. Columns (2-4) provide estimates for the effect of feedback on performance in tasks carried out thereafter. For all follow-up tasks, we find statistically insignificant effects. This indicates that the favorable short-term effect of positive feedback does not carry on to future tasks. Negative feedback has no impact, neither on subsequent nor future tasks.

Table 4: A non-persistent effect of feedback on performance

Performance	(1)	(2)	(3)	(4)
Positive feedback	0.115***	-0.010	0.073	-0.062
	(0.032)	(0.047)	(0.050)	(0.061)
Negative feedback	0.001	-0.049	0.023	-0.079
	(0.029)	(0.036)	(0.046)	(0.056)
Periods after feedback:	1	2	3	4
N	13075	10130	7350	4512

Notes: Linear Regression on future outcomes. Diving data. Specifications as in column (3) in Table 2. Standard errors are clustered on the individual level. *, **, and *** represents statistical significance at the 10 %, 5 %, and 1 % level, respectively.

5.4 Spillover effects on related tasks

Previously presented evidence shows the favorable effects of positive feedback on the task for which the feedback was obtained. In practice, individuals might do several tasks simultaneously, or a task containing different elements, that potentially influence each other. For example, Hecht, Tafkov, and Towry (2012) show spillover effects of incentive schemes in one task on a related, simultaneously conducted second task. Our settings allow us to study, both, a single-task and a multi-task environment.

Panel A presents the results for the single-task setup. As presented previously in Table 2, we find a performance-enhancing impact of positive feedback and no impact of negative feedback on performance. The difficulty is fixed ex-ante. That we find no impact of feedback on the difficulty can be regarded as a placebo outcome test and supports our identification strategy. Difficulty and performance evaluation jointly determine the combined outcome. Consequently, we observe a favorable effect of positive feedback on the total score.

Panel B exhibits the results for the multi-task environment. We observe favorable spillover effects. Receiving positive feedback in Task 1 enhances subsequent performance in Task 1 and the related Task 2. Negative feedback has no impact on either of the tasks. In the setup, performance in Task 1 and Task 2 are the most important determinants of combined success and the only ones that can be influenced by the task taker. Consistently, we also find a favorable influence of positive feedback on the total score.

Table 5: Spillover effects

Panel A: One isolated task, diving

	Task 1: Performance	Multiplier: Difficulty	Combined: Total score
Positive Feedback	0.115***	-0.002	1.071***
	(0.032)	(0.006)	(0.324)
Negative Feedback	0.001	-0.001	-0.029
	(0.029)	(0.005)	(0.282)

Panel B: Two simultaneous tasks, ski jumping

	Task 1: Performance	Task 2: Distance points	Combined: Total score
Positive Feedback	0.145***	1.692***	2.126***
	(0.034)	(0.634)	(0.693)
Negative Feedback	-0.049	0.072	-0.080
	(0.037)	(0.545)	(0.631)

Notes: Linear Regression estimates. Control variables as in column (3) in Table 2. Feedback was given previously for Task 1 only. Standard errors are clustered on the individual level. *, ** , and *** represents statistical significance at the 10 %, 5 %, and 1 % level, respectively.

5.5 Robustness

To ensure that our results are robust to different specifications we conduct several supplementary analyses. First, we consider alternative specifications of our key variables. In a first regression, we take the mean of all (five or seven) judges' ratings, instead of the performance, i.e., the mean of the (after discarding the extreme ratings) remaining three ratings, as an alternative outcome variable. With the treatment, the second key variable is (additionally) constructed in two different ways. Instead of subtracting the jury's performance assessment from the most extreme (positive/negative) discarded rating we deduct (a) the lowest (highest) rating included in the jury's performance assessment from the lowest (highest) discarded rating (Deviation positive/negative^{+}) and (b) the jury's performance assessment from the mean of the two discarded highest or lowest ratings (Deviation positive/negative^{++}, in diving only). Table 11 presents the results for these alternative specifications and shows robust estimates. We conclude from this that the result does neither depend on the concrete choice of the treatment variable, nor on the selection of the outcome variable.

Second, we consider different choices with respect to the sample that is used for the investigation. Data cleaning might offer some leeway to researchers influencing results. Thus, we provide additional analyses in Table 11 using (a) the full sample without any data cleaning and (b) without excluding failed attempts (but excluding boundary values as described in Section 3.1). We find robust results for both supplementary analyses, indicating that our data-cleaning step does not drive the results.

Third, to prove that nationality bias is not responsible for the effect, i.e., judges favor their compatriots and potentially influence other judges on the panel, we re-estimate the results excluding all athletes with a compatriot judge in the panel. If the effect would be driven by these individuals the results might just be some mechanical effect. Though, the effect is also found for individuals not sharing nationality with a judge.

6 Managerial implications and conclusions

Giving feedback is one of the most important tasks of managers. On a typical workday, managers regularly provide feedback to their teams. Some of this feedback is subconscious, such as facial expressions or nodding as a sign of appreciation and approval. Other feedback can be formal and dictated by the institution, as is the case with appraisal interviews. It can be constructive and substantive. But it can also be purely motivational. Common examples would be phrases like "Good job!" or "You can do better!" embedded in the context of everyday conversations.

The crucial question is whether such motivational feedback, given consciously by managers, can serve the goal of increasing the future productivity of workers. For both valences of feedback, i.e. positive and negative feedback, this question is not trivial. The appreciation that positive feedback expresses can motivate but also cause employees to rest on their laurels. Negative feedback can spur on but it can also hurt and discourage.

Our causal analysis indicates that managers can use positive feedback to enhance productivity. Our results show a favorable impact of positive feedback on (subsequent) performance. The heterogeneity analysis indicates that this favorable effect of positive feedback can be found for feedback recipients coming from varying cultural backgrounds, for recipients of both male and female gender, and for relatively more and less experienced recipients. We find that the favorable effect of positive feedback is short-term, repeatable, and with potentially favorable spillover to related tasks. The favorable impact of positive feedback is robust to the setup in which the activity is performed and is more pronounced in highly relevant situations. All this makes us confident that giving positive motivational feedback is a performance-enhancing strategy.

Furthermore, we find no significant impact of negative feedback on performance. This null effect might explain why managers and other raters are often reluctant to give negative feedback (Fisher, 1979), a phenomenon termed as leniency bias (Cheng, Hui, & Cascio, 2017) or MUM-effect (Rosen & Tesser, 1970). While in other contexts the lack of negative ratings is decreasing efficiency (Cannon & Witherspoon, 2005; Bolton, Kusterer, & Mans, 2019; Keser & Späth, 2021), we report no need to give negative motivational feedback.

Despite the robustness of our results, we acknowledge some limitations of our approach. First, our sample consists of internationally competing athletes. While their level of professionalism and self-discipline might be comparable to those of employees in highly competitive work environments, top athletes are not representative of the general population. Second, we consider an environment in which individuals receive feedback from multiple, external sources. Again, this is more comparable to daily life at large and competitive companies than at small firms. Third, we analyze a domain in which feedback recipients directly benefit from improvements in their performance, while feedback providers do not. In other domains, raters might be more prone to willfully bias their feedback.

Therefore, we suggest that future research could contrast our results to environments, in which feedback providers benefit from an increased performance more than feedback recipients do. Employees in such environments might be prone to exploitation when employers use positive feedback as a substitute for more substantial improvements in the employees' well-being. Furthermore, future research could analyze the long-term effects of positive and negative feedback.

With this study, we contribute to the literature that provides guidelines for optimal feedback (Balcazar et al., 1985; Alvero et al., 2001; Sleiman et al., 2020). Our causal analysis shows that positive feedback is improving performance, while negative feedback has no effect.

References

Abeler, J., Falk, A., Goette, L., & Huffman, D. (2011). Reference points and effort provision. *American Economic Review*, *101*(2), 470–492. doi: 10.1257/aer.101.2.470

Alvero, A. M., Bucklin, B. R., & Austin, J. (2001). An objective review of the effectiveness and essential characteristics of performance feedback in organizational settings (1985-1998). *Journal of Organizational Behavior Management*, *21*(1), 3–29. doi: 10.1300/J075v21n01_02

Azmat, G., & Iriberri, N. (2010). The importance of relative performance feedback information: Evidence from a natural experiment using high school students. *Journal of Public Economics*, *94*(7-8), 435–452. doi: 10.1016/j.jpubeco.2010.04.001

Bailey, C., & Fletcher, C. (2002). The impact of multiple source feedback on management development: findings from a longitudinal study. *Journal of Organizational Behavior*, *23*(7), 853–867. doi: 10.1002/job.167

Balcazar, F., Hopkins, B. L., & Suarez, Y. (1985). A critical, objective review of performance feedback. *Journal of Organizational Behavior Management*, *7*(3-4), 65–89. doi: 10.1300/J075v07n03_05

Bandiera, O., Larcinese, V., & Rasul, I. (2015). Blissful ignorance? a natural experiment on the effect of feedback on students' performance. *Labour Economics*, *34*, 13–25. doi: 10.1016/j.labeco.2015.02.002

Barron, K. (2021). Belief updating: does the 'good-news, bad-news' asymmetry extend to purely financial domains? *Experimental Economics*, *24*(1), 31–58. doi: 10.1007/s10683-020-09653-z

Bear, J. B., Cushenbery, L., London, M., & Sherman, G. D. (2017). Performance feedback, power retention, and the gender gap in leadership. *The Leadership Quarterly*, *28*(6), 721–740.

Berlin, N., & Dargnies, M.-P. (2016). Gender differences in reactions to feedback and willingness to compete. *Journal of Economic Behavior & Organization*, *130*, 320–336. doi: 10.1016/j.jebo.2016.08.002

Bolton, G. E., Kusterer, D. J., & Mans, J. (2019). Inflated reputations: Uncertainty, leniency, and moral wiggle room in trader feedback systems. *Management Science*, *65*(11), 5371–5391. doi: 10.1287/mnsc.2018.3191

Breiman, L. (2001). Random forests. *Machine Learning*, *45*(1), 5–32. doi: 10.1023/A:1010933404324

Burgers, C., Eden, A., van Engelenburg, M. D., & Buningh, S. (2015). How feedback boosts motivation and play in a brain-training game. *Computers in Human Behavior*, *48*, 94–103. doi: 10.1016/j.chb.2015.01.038

Cannon, M. D., & Witherspoon, R. (2005). Actionable feedback: Unlocking the power of learning and performance improvement. *Academy of Management Perspectives*, *19*(2), 120–134. doi: 10.5465/ame.2005.16965107

Carpentier, J., & Mageau, G. A. (2013). When change-oriented feedback enhances motivation, well-being and performance: A look at autonomy-supportive feedback in sport. *Psychology of Sport and Exercise*, *14*(3), 423–435. doi: 10.1016/j.psychsport.2013.01.003

Cason, & Mui. (1998). Social influence in the sequential dictator game. *Journal of Mathematical Psychology*, *42*(2/3), 248–265. doi: 10.1006/jmps.1998.1213

Cheng, K. H. C., Hui, C. H., & Cascio, W. F. (2017). Leniency bias in performance ratings: The big-five correlates. *Frontiers in Psychology*, *8*, 521. doi: 10.3389/fpsyg.2017.00521

Choi, E., Johnson, D. A., Moon, K., & Oah, S. (2018). Effects of positive and negative feedback sequence on work performance and emotional responses. *Journal of Organizational Behavior Management*, *38*(2-3), 97–115. doi: 10.1080/01608061.2017.1423151

Coutts, A. (2019). Good news and bad news are still news: experimental evidence on belief updating. *Experimental Economics*, *22*(2), 369–395. doi: 10.1007/s10683-018-9572-5

Damisch, L., Mussweiler, T., & Plessner, H. (2006). Olympic medals as fruits of comparison? assimilation and contrast in sequential performance judgments. *Journal* *of Experimental Psychology: Applied*, *12*(3), 166.

Deci, E. L., & Casico, W. F. (1972). *Changes in intrinsic motivation as a function of negative feedback and threats.*

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

Eggers, J. P., & Suh, J.-H. (2019). Experience and behavior: How negative feedback in new versus experienced domains affects firm action and subsequent performance. *Academy of Management Journal*, *62*(2), 309–334. doi: 10.5465/amj.2017.0046

Eil, D., & Rao, J. M. (2011). The good news-bad news effect: Asymmetric processing of objective information about yourself. *American Economic Journal: Microeconomics*, *3*(2), 114–138. doi: 10.1257/mic.3.2.114

Ertac, S. (2011). Does self-relevance affect information processing? experimental evidence on the response to performance and non-performance feedback. *Journal of Economic Behavior & Organization*, *80*(3), 532–545. doi: 10.1016/j.jebo.2011.05.012

Findlay, L. C., & Ste-Marie, D. M. (2004). A reputation bias in figure skating judging. *Journal of Sport and Exercise Psychology*, *26*(1), 154–166. doi: 10.1123/jsep.26.1.154

Fisher, C. D. (1979). Transmission of positive and negative feedback to subordinates: A laboratory investigation. *Journal of Applied Psychology*, *64*(5), 533. doi: 10.1037/0021-9010.64.5.533

Fong, C. J., Patall, E. A., Vasquez, A. C., & Stautberg, S. (2019). A meta-analysis of negative feedback on intrinsic motivation. *Educational Psychology Review*, *31*(1), 121–162. doi: 10.1007/s10648-018-9446-6

Ginsburgh, V. A., & Van Ours, J. C. (2003). Expert opinion and compensation: Evidence from a musical competition. *The American Economic Review*, *93*(1), 289–296. doi: 10.1257/000282803321455296

Goller, D., & Heiniger, S. (2022). A general framework to quantify the event importance in multi-event contests. *arXiv preprint arXiv:2207.02316*.

Goller, D., & Krumer, A. (2020). Let's meet as usual: Do games played on non-frequent days differ? evidence from top european soccer leagues. *European Journal of Operational Research*, *286*(2), 740–754. doi: 10.1016/j.ejor.2020.03.062

Harrison, S. H., & Rouse, E. D. (2015). An inductive study of feedback interactions over the course of creative projects. *Academy of Management Journal*, *58*(2), 375–404.

Hattie, J., & Timperley, H. (2007). The power of feedback. *Review of educational research*, *77*(1), 81–112.

Hecht, G., Tafkov, I., & Towry, K. L. (2012). Performance spillover in a multitask environment. *Contemporary Accounting Research*, *29*(2), 563–589. doi: 10.1111/j.1911-3846.2011.01114.x

Heiniger, S., & Mercier, H. (2021). Judging the judges: evaluating the accuracy and national bias of international gymnastics judges. *Journal of Quantitative Analysis in Sports*, *17*(4), 289–305. doi: 10.1515/jqas-2019-0113

Henrich, J., Heine, S. J., & Norenzayan, A. (2010). Most people are not weird. *Nature*, *466*(7302), 29. doi: 10.1038/466029a

Hofstede, G. (2011). Dimensionalizing cultures: The hofstede model in context. *Online Readings in Psychology and Culture*, *2*(1). doi: 10.9707/2307-0919.1014

Itzchakov, G., & Latham, G. P. (2020). The moderating effect of performance feedback and the mediating effect of self–set goals on the primed goal–performance relationship. *Applied Psychology*, *69*(2), 379–414. doi: 10.1111/apps.12176

Johnson, D. A. (2013). A component analysis of the impact of evaluative and objective feedback on performance. *Journal of Organizational Behavior Management*, *33*(2), 89–103. doi: 10.1080/01608061.2013.785879

Kennedy, E. H., Ma, Z., McHugh, M. D., & Small, D. S. (2017). Non-parametric methods for doubly robust estimation of continuous treatment effects. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *79*(4), 1229–1245. doi: 10.1111/rssb.12212

Keser, C., & Späth, M. (2021). The value of bad ratings: An experiment on the impact of distortions in reputation systems. *Journal of Behavioral and Experimental*

Economics, *95*, 101782. doi: 10.1016/j.socec.2021.101782

Kim, Y. J., & Kim, J. (2020). Does negative feedback benefit (or harm) recipient creativity? the role of the direction of feedback flow. *Academy of Management Journal*, *63*(2), 584–612. doi: 10.5465/amj.2016.1196

Kluger, A. N., & DeNisi, A. (1996). The effects of feedback interventions on performance: A historical review, a meta-analysis, and a preliminary feedback intervention theory. *Psychological Bulletin*, *119*(2), 254–284. doi: 10.1037/0033-2909.119.2.254

Krumer, A., Otto, F., & Pawlowski, T. (2022). Nationalistic bias among international experts: Evidence from professional ski jumping. *The Scandinavian Journal of Economics*, *124*(1), 278–300. doi: 10.1111/sjoe.12451

Kuzmanovic, B., Jefferson, A., & Vogeley, K. (2015). Self-specific optimism bias in belief updating is associated with high trait optimism. *Journal of Behavioral Decision Making*, *28*(3), 281–293. doi: 10.1002/bdm.1849

Lee, J., Lee, J. M., & Kim, J.-Y. (2021). The role of attribution in learning from performance feedback: Behavioral perspective on the choice between alliances and acquisitions. *Academy of Management Journal*((In-press)).

Markus, H. R., & Kitayama, S. (1991). Culture and the self: Implications for cognition, emotion, and motivation. *Psychological Review*, *98*(2), 224. doi: 10.1037/0033-295X.98.2.224

Möbius, M. M., Niederle, M., Niehaus, P., & Rosenblat, T. S. (2022). Managing self-confidence: Theory and experimental evidence. *Management Science*. doi: 10.1287/mnsc.2021.4294

Moore, D. A., & Healy, P. J. (2008). The trouble with overconfidence. *Psychological Review*, *115*(2), 502. doi: 10.1037/0033-295X.115.2.502

Morgan, H. N., & Rotthoff, K. W. (2014). The harder the task, the higher the score: Findings of a difficulty bias. *Economic Inquiry*, *52*(3), 1014–1026. doi: 10.1111/ecin.12074

Morgeson, F. P., DeRue, D. S., & Karam, E. P. (2010). Leadership in teams: A functional approach to understanding leadership structures and processes. *Journal of*

Management, *36*(1), 5–39. doi: 10.1177/0149206309347376

Muthukrishna, M., Bell, A. V., Henrich, J., Curtin, C. M., Gedranovich, A., McInerney, J., & Thue, B. (2020). Beyond western, educated, industrial, rich, and democratic (weird) psychology: Measuring and mapping scales of cultural and psychological distance. *Psychological science*, *31*(6), 678–701. doi: 10.1177/0956797620916782

Page, K., & Page, L. (2010). Alone against the crowd: Individual differences in referees' ability to cope under pressure. *Journal of Economic Psychology*, *31*(2), 192–199. doi: 10.1016/j.joep.2009.08.007

Podsakoff, P. M., & Farh, J.-L. (1989). Effects of feedback sign and credibility on goal setting and task performance. *Organizational Behavior and Human Decision Processes*, *44*(1), 45–67. doi: 10.1016/0749-5978(89)90034-4

Pulford, B. D., & Colman, A. M. (1997). Overconfidence: Feedback and item difficulty effects. *Personality and Individual Differences*, *23*(1), 125–133. doi: 10.1016/S0191-8869(97)00028-7

Rhee, M., Alexandra, V., & Powell, K. S. (2020). Individualism-collectivism cultural differences in performance feedback theory. *Cross Cultural & Strategic Management*, *27*(3), 343–364. doi: 10.1108/CCSM-05-2019-0100

Roberts, T.-A., & Nolen-Hoeksema, S. (1994). Gender comparisons in responsiveness to others' evaluations in achievement settings. *Psychology of Women Quarterly*, *18*(2), 221–240. doi: 10.1111/j.1471-6402.1994.tb00452.x

Rosen, S., & Tesser, A. (1970). On reluctance to communicate undesirable information: The mum effect. *Sociometry*, *33*(3), 253. doi: 10.2307/2786156

Sandberg, A. (2018). Competing identities: a field study of in-group bias among professional evaluators. *The Economic Journal*, *128*(613), 2131–2159. doi: 10.1111/ecoj.12513

Sharot, T., Kanai, R., Marston, D., Korn, C. W., Rees, G., & Dolan, R. J. (2012). Selectively altering belief formation in the human brain. *Proceedings of the National Academy of Sciences of the United States of America*, *109*(42), 17058–17062. doi: 10.1073/pnas.1205828109

Sleiman, A. A., Sigurjonsdottir, S., Elnes, A., Gage, N. A., & Gravina, N. E. (2020). A quantitative review of performance feedback in organizational settings (1998-2018). *Journal of Organizational Behavior Management*, *40*(3-4), 303–332. doi: 10.1080/01608061.2020.1823300

Smither, J. W., London, M., & Reilley, R. R. (2005). Does performance improve following multisource feedback? a theoretical model, meta-analysis, and review of empirical findings. *Personnel Psychology*, *58*(1), 33–66. doi: 10.1111/j.1744-6570.2005.514_1.x

Stone, E. F., & Stone, D. L. (1984). The effects of multiple sources of performance feedback and feedback favorability on self-perceived task competence and perceived feedback accuracy. *Journal of Management*, *10*(3), 371–378. doi: 10.1177/014920638401000311

Sully De Luque, M. F., & Sommer, S. M. (2000). The impact of culture on feedback-seeking behavior: An integrated model and propositions. *Academy of Management Review*, *25*(4), 829–849. doi: 10.5465/amr.2000.3707736

Vancouver, J. B., & Tischner, E. C. (2004). The effect of feedback sign on task performance depends on self-concept discrepancies. *The Journal of Applied Psychology*, *89*(6), 1092–1098. doi: 10.1037/0021-9010.89.6.1092

Villeval, M. C. (2020). Performance feedback and peer effects. In K. F. Zimmermann (Ed.), *Handbook of labor, human resources and population economics* (pp. 1–38). Cham: Springer International Publishing. doi: 10.1007/978-3-319-57365-6_126-1

Waldersee, R., & Luthans, F. (1994). The impact of positive and corrective feedback on customer service performance. *Journal of Organizational Behavior*, *15*(1), 83–95. doi: 10.1002/job.4030150109

Zitzewitz, E. (2006). Nationalism in winter sports judging and its lessons for organizational decision making. *Journal of Economics & Management Strategy*, *15*(1), 67–99.

Appendix

Descriptive statistics

Table 6: Full descriptive statistics

	Diving		Ski jumping	
	Mean	Std. dev.	Mean	Std. dev.
Treatments:				
Positive feedback (deviation positive)	0.426	(0.286)	0.316	(0.262)
Negative feedback (deviation negative)	0.477	(0.320)	0.357	(0.290)
Positive feedback$^+$	0.314	(0.297)	0.179	(0.258)
Negative feedback$^+$	0.363	(0.328)	0.218	(0.289)
Future positive feedback	0.439	(0.301)		
Future negative feedback	0.489	(0.325)		
Outcomes:				
Performance (rem. 3 judges' ratings)	7.119	(1.189)	17.771	(0.744)
Performance (all 5 / 7 judges' ratings)	7.110	(1.182)	17.765	(0.741)
Score	68.737	(14.557)	118.647	(16.204)
Distance			122.608	(11.837)
Covariates:				
Difficulty	3.211	(0.331)		
Compatriot judge	0.248		0.457	
Home event	0.099		0.127	
Final	0.291			
Female	0.450			
Age	22.429	(3.789)	26.836	(4.949)
Current ranking	8.490	(9.655)	15.357	(8.582)
Start order	9.490	(11.082)		
Points behind leader	31.491	(31.011)	19.247	(10.132)
In range (within 5 pts. to threshold)	0.264			
Gate points			0.093	(3.270)
Wind points			-0.291	(8.225)
Prev. performance	7.270	(0.958)	17.854	(0.580)
Prev. SD performance	0.130	(0.151)	0.157	(0.159)
Prev. wind points			-1.685	(8.136)
Prev. gate points			-0.163	(4.386)
Prev. distance			123.940	(11.143)
Prev. difficulty	3.166	(0.317)		
N		13075		4529

Notes: Mean and standard deviation (in parentheses; for non-binary variables). Some variables only observed in one of the data sets. $^+$Alternative definition as defined in the main text.

Placebo and balancing tests

Table 7: Placebo treatment regressions

	Judges' ratings 3	Judges' ratings 5	Judges' ratings 7	Score
Future positive feedback	0.028	0.030	0.027	-0.012
	(0.035)	(0.035)	(0.035)	(0.345)
Future negative feedback	-0.045	-0.043	-0.041	-0.437
	(0.035)	(0.035)	(0.035)	(0.318)
N	10256	10256	10256	10256

Notes: Linear Regression on the outcome mentioned in the column header. 3, 5, and 7 refer to discarding four, two, or none of the extreme judges' ratings. Diving data. Pseudo-treatment is the deviation of next (future) jump. Jumps 2–4/5 only. Specifications as in column (3) in Table 2. Standard errors are clustered on the individual level. *, **, and *** represents statistical significance at the 10 %, 5 %, and 1 % level, respectively.

Table 8: Balancing Tests

	Diving					
	Compatriot judge		Home event		SD prev. perform.	
	(1)	(2)	(3)	(4)	(5)	(6)
Feedback positive	-0.000		-0.004		0.007	
	(0.014)		(0.008)		(0.005)	
Feedback negative		-0.017		-0.012		0.002
		(0.012)		(0.007)		(0.005)
	Difficulty		Final			
	(1)	(2)	(3)	(4)		
Feedback positive	-0.005		-0.018			
	(0.007)		(0.013)			
Feedback negative		0.006		-0.017		
		(0.006)		(0.013)		
	Ski jumping					
	Compatriot judge		Home event		Prev. distance	
	(1)	(2)	(3)	(4)	(5)	(6)
Feedback positive	-0.045		-0.023		0.918	
	(0.028)		(0.018)		(0.696)	
Feedback negative		-0.010		-0.048***		0.345
		(0.022)		(0.017)		(0.674)
	Prev. gate		SD prev. perform.			
	(1)	(2)	(3)	(4)		
Feedback positive	-0.304		0.024			
	(0.306)		(0.092)			
Feedback negative		0.031		0.018		
		(0.217)		(0.095)		

Notes: Linear Regression estimates. Each regression includes athlete fixed-effects. Standard errors are clustered on the individual level. *, **, and *** represents statistical significance at the 10 %, 5 %, and 1 %, respectively.

Additional and full results tables

Table 9: Feedback on performance – sensitivity to different specifications, ski jumping

	Ski jumping			
Performance	(1)	(2)	(3)	(4)
Positive feedback	0.201***	0.180***	0.145***	0.107***
	(0.035)	(0.036)	(0.034)	(0.034)
Negative feedback	-0.063	-0.055	-0.049	-0.026
	(0.043)	(0.041)	(0.037)	(0.041)
Prev. jury assessment	0.593***	0.465***	0.402***	0.329***
	(0.027)	(0.041)	(0.043)	(0.031)
Prev. wind points		0.044***	0.040***	0.036***
		(0.002)	(0.002)	(0.002)
Prev. gate points		0.003	0.002	0.000
		(0.003)	(0.003)	(0.003)
Prev. distance		0.002***	0.003***	0.002**
		(0.001)	(0.001)	(0.002)
Wind points		-0.041***	-0.038***	-0.036***
		(0.002)	(0.002)	(0.002)
Gate points		-0.020***	-0.019***	-0.019***
		(0.003)	(0.002)	(0.003)
Points behind		-0.015***	-0.016***	-0.015***
		(0.002)	(0.002)	(0.002)
Compatriot judge		0.021	0.016	0.024
		(0.020)	(0.022)	(0.023)
Home event		0.013	0.028	0.041
		(0.032)	(0.032)	(0.035)
Start order		0.002	-0.003	-0.005*
		(0.002)	(0.002)	(0.003)
SD prev. judges' ratings.		-0.019	-0.001	-0.017
		(0.061)	(0.063)	(0.065)
Athlete Fixed Effect			x	
Athlete x Season FE				x
N	4529	4529	4529	4529

Notes: Linear regression. Prev. (= previous) refers to a lagged variable from the previous jump. SD = standard deviation. Standard errors are clustered on the individual level. *, **, and *** represents statistical significance at the 10 %, 5 %, and 1 % level, respectively.

Table 10: Feedback on performance – sensitivity to different specifications, diving

	Diving			
Performance	(1)	(2)	(3)	(4)
Positive Feedback	0.242***	0.208***	0.115***	0.100***
	(0.036)	(0.034)	(0.032)	(0.035)
Negative Feedback	0.018	0.024	0.001	0.007
	(0.030)	(0.030)	(0.029)	(0.030)
Prev. jury assessment	0.430***	0.284***	0.103***	0.073***
	(0.026)	(0.022)	(0.016)	(0.016)
Prev. difficulty	0.794***	0.540***	0.147	0.228**
	(0.079)	(0.087)	(0.091)	(0.100)
SD prev. judges' ratings		0.095	0.056	0.029
		(0.067)	(0.067)	(0.070)
Compatriot judge		-0.015	-0.024	-0.016
		(0.024)	(0.022)	(0.025)
Home event		0.129***	0.164***	0.196***
		(0.038)	(0.045)	(0.054)
Current ranking		-0.020***	0.000	0.011***
		(0.002)	(0.002)	(0.003)
Start order		-0.003***	-0.006***	-0.009***
		(0.001)	(0.001)	(0.001)
Points behind		-0.003***	-0.000	0.001
		(0.001)	(0.000)	(0.001)
Penalty		-0.288	-0.362*	-0.310
		(0.187)	(0.187)	(0.200)
Jump and Event Fixed Effect		x	x	x
Athlete Fixed Effect			x	
Athlete x Season Fixed Effects				x
N	13075	13075	13075	13075

Notes: Prev. (= previous) refers to a lagged variable from the previous jump. SD = standard deviation. Fixed effects for *Events* are 1m and 3m Springboard and 10m Platform, and the five (female) or six (male) jumps. Standard errors are clustered on the individual level. *, **, and *** represents statistical significance at the 10 %, 5 %, and 1 %, respectively.

Table 11: Robustness Checks

	Ski jumping		Diving	
	Positive	Negative	Positive	Negative
	Feedback		Feedback	
Baseline results	0.121***	-0.036	0.115***	0.001
	(0.033)	(0.044)	(0.032)	(0.029)
Other outcome variable	0.129***	-0.070*	0.111***	0.005
(all ratings, incl. discarded)	(0.035)	(0.038)	(0.032)	(0.029)
Treatment definition 2	0.119***	-0.062	0.109***	0.005
(Discarded vs. last credited)	(0.033)	(0.039)	(0.032)	(0.030)
Treatment definition 3			0.127***	0.007
(Mean discarded vs. mean credited)			(0.043)	(0.039)
Without data cleaning	0.124***	-0.056	0.059*	0.017
	(0.044)	(0.047)	(0.035)	(0.031)
Without dropping failed attempts	0.124***	-0.042	0.109***	0.031
	(0.045)	(0.042)	(0.037)	(0.033)
Only athletes not sharing	0.175***	-0.044	0.113***	-0.021
nationality with a judge	(0.040)	(0.053)	(0.038)	(0.034)
Only jumps with no variance	0.132***	-0.044	0.143***	0.056
in scoring ratings	(0.043)	(0.051)	(0.043)	(0.038)

Notes: Linear regression. Every line represents two separate regressions, one in each data set. Specification as in column (3) in Table 2. Standard errors are clustered on the individual level. *, **, and *** represents statistical significance at the 10 %, 5 %, and 1 %, respectively.