Feedback quality and performance in organisations

Drouvelis, Michalis; Paiardini, Paola

DOI:
10.1016/j.leaqua.2021.101534

Document Version
Peer reviewed version

Citation for published version (Harvard):
Drouvelis, M & Paiardini, P 2021, 'Feedback quality and performance in organisations', *The Leadership Quarterly*. https://doi.org/10.1016/j.leaqua.2021.101534

Feedback quality and performance in organisations

Michalis Drouvelis[*, **]

University of Birmingham & CESifo

Paola Paiardini[***]

University of Birmingham

May 2021

Abstract

The provision of feedback is fundamental for promoting employee performance in modern organisations; however, little is known about how the quality of feedback affects performance. We report an experiment where subjects perform a real-effort task repeatedly in a flat-wage environment which varies the quality of feedback across treatments. In the baseline treatment, subjects receive no feedback about their rank in their group. In the two main treatments, feedback quality varies in that subjects know ("High-Quality Feedback") or do not know ("Low-Quality Feedback") their exact rank in their group. We show that the quality of feedback is an important driver of productivity. Average performance is significantly higher for high-quality feedback than for low-quality or no feedback, where no significant overall performance differences are observed. Our results have significant implications for designing and implementing cost-effective policies within organisations.

Keywords: Feedback quality; performance; non-monetary incentives; laboratory experiment.
JEL Classification: C91; D01; J30.

Acknowledgements: We thank John Antonakis, the Method Advisor, two anonymous reviewers, and Ghazala Azmat, David Gill, Alessandra Guariglia, Werner Güth, Nagore Iriberri, Kimberley Scharf, Marie Claire Villeval and participants at the BI Norwegian Business School, the VATT Institute for Economic Research, the 2018 Birmingham Experimental Economics Workshop on Social Preferences, the 2018 Foundations of Utility and Risk (FUR) Conference in York, the 2018 Tax-Day at the Max Plank Institute in Munich, and the GATE seminar at the University of Lyon for their useful comments. We thank Alexandru Blinda for his excellent programming assistance. Financial support from the University of Birmingham is gratefully acknowledged.

[*] Department of Economics, University of Birmingham, Edgbaston, B15 2TT. Email: m.drouvelis@bham.ac.uk.
[**] CESifo, Munich, Germany.
[***] Department of Economics, University of Birmingham, Edgbaston, B15 2TT. Email: p.paiardini@bham.ac.uk.

1. Introduction

Modern organisations employ a broad range of incentive schemes to promote employee performance and career prospects (see Prendergast, 1999). Providing feedback to encourage employees to improve their performance is a common management practice in workplace environments (see Murphy & Cleveland, 1991). Existing research has also demonstrated that the use of performance appraisals is fundamental in motivating workers (for an overview, see Prewitt, 2007). While the goal of these appraisals is to enhance performance, conventional wisdom holds that for feedback to be effective, it should be precise. In practice, however, as Longenecker et al. (1987) note, managers are less concerned with the accuracy of performance appraisals and, instead, most of their attention is particularly directed at how to financially motivate and reward their subordinates. It therefore naturally follows to ask: Does the lack of quality when providing feedback affect performance in organisations, and if so, how?

Further motivation for our research question stems from anecdotal evidence highlighting that the quality of feedback has adverse effects on employees' motivation to work hard. Importantly, these effects have not yet been measured. A recent Forbes article (Jackson, 2012), entitled "Ten Biggest Mistakes Bosses Make in Performance Reviews", lists as the top mistake in performance review meetings that the feedback provided to employees is "too vague".[1] In particular, it is often reported that employees receive vague feedback on their performance, such as "You are doing a good job" or "Keep it up". The same article highlights that, in many instances, annual performance review meetings give no specific feedback about the work performed, and employees think that everything is perfect until they are fired that they realise it is not. The role of vague feedback has also been recognised by Gibbs (1991, p. 10) who writes that "If a runner is far out in front, his motivation to run hard is smaller than otherwise. Running fast is intrinsically rewarding, but it is also exhausting and risks injury, so such a runner is likely to slack off somewhat. Similarly, a runner who is far behind the lead also has little incentive to run hard. Those with the greatest reason to run as fast as they can are those who are in a tight race for the finish". In contrast, the use of high-quality feedback (e.g., stack rankings) may be one potential component of management strategies that can enhance productivity. In this paper, we seek to address how varying the quality of feedback causally affects performance (when monetary incentives are ruled out).

[1] For more details, see https://www.forbes.com/sites/ericjackson/2012/01/09/ten-reasons-performance-reviews-are-done-terribly/#37cb2cf25ee0.

This paper aims to broaden the existing interdisciplinary literature on feedback interventions by causally analysing how managers can spur employees' effort using feedback as a motivation tool. A considerable body of evidence in behavioural management and applied psychology examines how feedback interventions affect performance (Kluger & DeNisi, 1996). In particular, several theories (such as the control theory (see Podsakoff & Farh, 1989), goal-setting theory (see Locke & Latham, 1990) and social cognition theory (see Bandura, 1991)) suggest that feedback is a key component for shaping task performance. Past meta-analyses demonstrate that feedback can improve performance; this was done using constructs like "feedback" per se (Stajkovic & Luthans, 2003) and "contingent rewards" – which are not economic rewards but essentially provide feedback on whether performance targets have been achieved – and by giving constructive or corrective feedback (Judge & Piccolo, 2004; Lowe et al. 1996).[2] For example, Podsakoff et al. (1990; 1996) and Podsakoff et al. (1990) include questionnaire items which, in addition to feedback, acknowledge that the measured style of leadership has to do with recognition too. Relatedly, the Bass–Avolio model (Avolio et al., 1995; Bass & Avolio, 1997) is concerned with communicating goals and then expressing satisfaction when these goals are reached.[3]

Yet, there are two major issues with regard to the above approaches and research theories. First, some of the elicited measures confound feedback and rewards/recognition, thus not allowing a clean evaluation solely of the feedback effects. Second, and to the best of our knowledge, in the management field there have not been any exogenous manipulations of feedback per se (for an exception in a field setting, see Peterson & Luthans, 2006) and oftentimes manipulations are based on hypothetical situations (vignettes), which are rather prone to demand effects (see Lonati et al., 2018).[4] In summary, the findings from these studies are usually hypothetical and subject to demand effects (Rosenthal & Rosnow, 2009; Zizzo, 2010).

While the existing evidence from the field of behavioural management and personal psychology largely claims that feedback matters (Kluger & DeNisi, 1996),[5] very little of this

[2] Items from this questionnaire include the following: "Always gives me positive feedback when I perform well; gives me special recognition when my work is very good; commends me when I do a better than average job; personally compliments me when I do outstanding work; frequently does not acknowledge my good performance".
[3] For example, two of their questionnaire items are as follows: "Makes clear what one can expect to receive when performance goals are achieved; expresses satisfaction when I meet expectations".
[4] In the management field, it is rarely the case that real effort-incentivised tasks are examined (for exceptions, see Meslec et al., 2020; Steffens et al., 2018).
[5] The role of feedback provision has also been explored beyond social sciences such as in medical contexts (see Winickoff et al., 1984).

research incorporates economic perspectives and methods into leadership research. Surprisingly, there has been little research interaction until now between the fields of leadership and economics although they do share common goals and objectives (Zehnder et al., 2017). Our study also contributes towards filling this gap in the literature by taking advantage of the benefits of the experimental economics methodology to inform research in the leadership field. We implement a laboratory experiment, the design of which exogenously varies the type of feedback that individuals receive when performing a monetarily incentivised real-effort task. Unlike previous management studies, we study an environment in which the provision of feedback is not confounded by other factors (such as rewards or social recognition) which may also act as reinforcers on task performance. Our design allows us to draw causal conclusions about the impact of pure feedback on performance outcomes, ruling out potential endogeneity issues that might confound the interpretation of our findings, making them policy-relevant (see Antonakis et al., 2016). Put simply, in our experiment the direction of the relationship between feedback and outcomes can be determined such that it can be shown that our outcome variable (task performance) is, indeed, caused by the type of feedback provided. Identifying a clear causal link is of particular importance because it differentiates us from how leadership and management scholars (using questionnaire measures) often study the relationship between two variables of interest (for a discussion, see Garretsen et al., 2020), an approach that could suffer from endogeneity bias. We come back and discuss this point in our concluding section.

The role of feedback has attracted much interest among applied psychologists, enabling them to propose several theories and research paradigms that contain the concept of feedback as a key component (see Kluger & DeNisi, 1996, for a discussion). Closer to our setting is the feedback intervention theory, positing that feedback interventions may influence task-motivation and task-learning processes and feedback cues may shift attention away from the task-relevant aspects. Changing the locus of attention from the task activity to the non-task aspects of the intervention (meta-task processes) involves the reallocation of cognitive resources and attenuates performance due to the depletion of cognitive resources. The fewer cognitive resources needed for task performance, the more positive the effects of a feedback intervention will be. In practice, certain feedback cues may direct attention to task-motivation and learning processes and thus improve task performance, whereas other feedback messages may shift attention to meta-task processes, leading to performance loss. For instance, in a study by Butler (1987) task-focused feedback interventions (i.e., giving specific comments on students' coursework) increased task engagement and performance; however, giving only grades did not influence performance as it promoted ego involvement. We use the feedback

intervention theory as our conceptual framework to organise the discussion and formulation of our hypotheses in relation to our context.

In addition, we broaden the economics literature by shedding light on the causal impact of feedback quality on performance using a laboratory experiment in which individuals perform a real-effort task under a flat-wage environment (see, for example, Falk & Ichino, 2006; Charness et al., 2014; Azmat & Iriberri, 2016). The rationale for analysing behaviour in a flat-wage environment is that it allows us to focus on the pure effects of feedback quality on performance, ruling out any potential behavioural confounds that might stem from competing for financial resources. Beyond the methodological importance of our design, our experimental setting resembles several real-life working environments where remuneration schemes are not tied to performance. This is frequently the case in the public sector, but flat-wage schemes are also observed in firms (e.g., Bartling & von Siemens, 2010). Holmström and Milgrom (1991) also note that "employment contracts so often specify fixed wages and more generally that incentives within firms appear to be so muted, especially compared to those of the market". Consequently, even in situations where employees receive flat wages for their work, they are still evaluated by their manager and such performance reviews typically form the basis for promotion opportunities as well as employment duration.

Our research is closest to studies looking at how the provision of feedback causally affects performance in flat-wage settings.[6] The main message from the existing literature is that offering feedback on relative performance and rank leads to higher performance (for an exception, see Barankay, 2012). For example, Charness et al. (2014) examine whether subjects are willing to alter their performance and, thus, their rank position, either by sabotaging others' performance or by buying extra output. Their findings indicate that offering relative rank feedback increases output with regard to the baseline no-feedback treatment, and subjects are willing to engage in costly sabotage and cheating activities to improve their relative rank, thus offsetting the positive effects of relative rank feedback. Relatedly, Tafkov (2013) also finds that relative performance feedback has positive effects on the overall output in a setting where compensation is not tied to performance. Kuhnen and Tymula (2012) design a within-subject experiment, where subjects perform a multiplication task in a flat-wage environment, randomly varying the likelihood of receiving relative performance feedback ("no feedback", "likely

[6] The performance-enhancing effects of relative performance/rank feedback have also been documented in settings where remuneration is tied to performance (e.g., Blanes i Vidal & Nossol, 2011; Azmat & Iriberri, 2010), but this relationship may also depend on the incentive scheme (e.g., Hannan et al., 2008; Eriksson et al., 2009). For a theoretical discussion of the effects of relative performance information, see Hölmstrom (1979) and Lizzeri et al. (2002).

feedback", and "sure feedback") across periods. They find that relative performance feedback can be strategically used to improve employee performance in organisations and, in particular, subjects who ranked better (worse) than expected decrease (increase) their productivity and expect a better (worse) rank in future interactions. This pattern suggests that feedback may generate a ratcheting effect in effort provision. Additionally, Gerhards and Siemer (2016) study the effects of private and public feedback, always in a flat-wage environment. In the private feedback treatment, the best performer within a group is privately announced, whereas in the public feedback treatment, each group's best-performing subject is additionally publicly recognised. These treatments are compared with no-feedback treatment as the baseline. The main findings from their study show that the provision of relative feedback significantly increases productivity compared to the no-feedback treatment; however, there is little effect on performance with regard to the two main feedback treatments. More recently, Gill et al. (2019) analyse the impact of rank-order relative-performance feedback in the context of both a numerical and verbal real-effort task. They find that participants react to the specific rank that they receive: subjects work harder after being ranked first or last. We also contribute to this strand of literature by measuring the effects of varying the quality of the feedback content on productivity.

A study related to our paper is that of Hannan et al. (2008), who look at the precision of feedback by distinguishing between two levels of feedback: "coarse" and "fine". Under coarse feedback, subjects can assess whether they are above or below the mean, but not the specific percentile of their relative performance. Their treatments consider a piece rate and a tournament incentive scheme. Their findings show asymmetric effects on performance: providing relative performance feedback deteriorates the mean performance of participants compensated under a tournament incentive scheme, but only in the fine feedback treatment; however, providing relative performance feedback improves the mean performance of participants compensated under an individual incentive scheme regardless of the feedback precision. It is worth noting that their setting is distinct from ours, as the observed effects in Hannan et al.'s experiment are confounded by monetary motives that are likely to emerge from competing for resources (unlike our experiment where such concerns have been ruled out). In addition, our experiment focuses on the quality of the feedback content provided and keeps the feedback precision (i.e., in which percentile subjects are) constant across treatments.

To understand the effects of feedback quality on task performance, we employ a between-subjects design, consisting of three treatments. In our main treatments, subjects

receive feedback about their own performance rank within their group, which can be specific to their rank – i.e., subjects know their exact rank in their group (high-quality treatment). In contrast, in the low-quality treatment, subjects receive feedback messages that are not informative of their rank in their group. As a baseline treatment, we include a treatment where no feedback is provided.

Our main findings are summarised as follows. We replicate results from previous studies (see Charness et al., 2014; Kuhnen & Tymula, 2012), demonstrating that even with fixed compensation, subjects exert positive effort. Crucially, this positive effort effect depends on the quality of the feedback that subjects receive. In particular, when subjects receive high-quality feedback about their own rank among their peers, average performance is significantly higher compared to receiving either low-quality or no feedback at all. Specifically, receiving low-quality feedback increases performance levels relative to the no-feedback treatment, but not significantly so. Overall, these findings provide useful insights for the public sector and private organisations that typically use relative performance feedback as a performance-enhancing mechanism. We offer new evidence that for feedback to be effective in raising productivity, it has to be of high quality. Our experiment broadens the existing interdisciplinary literature by highlighting the causal role of the quality of non-monetary incentives, such as feedback that leaders and managers can use to spur on employees' performance (see also Sloof & von Siemens, 2019).

Our findings are organised by the feedback intervention theory proposed by Kluger and DeNisi (1996). Relying on this framework, high-quality feedback cues direct attention to task-learning and task-motivation processes and have positive effects on performance. The provision of high-quality feedback cues requires fewer cognitive resources for task performance, which in turn has more positive effects on performance. In contrast, low-quality feedback shifts the focus of attention to meta-task processes, decreasing task involvement and debilitating performance.

More specifically, our results indicate that in organisational environments where extrinsic rewards for performance are non-existent, free-riding incentives can be mitigated by providing high-quality feedback to workers regarding their rank in the group. In light of these findings, it is possible that by paying special attention to promoting quality in performance appraisals, managers and employers can benefit using a cost-effective incentive mechanism that raises firms' productivity levels. Our paper further widens the expansive literature in economics, human resource management, and psychology, showing that non-financial

incentives can influence behaviour in various decision-making environments (see Etzioni, 1971; Gneezy & Rustichini, 2000; Cropanzano & Mitchell, 2005; Kosfeld & Neckermann, 2011; Eriksson & Villeval, 2012). Unlike existing studies in management and applied psychology fields, which typically use questionnaire techniques (suffering from potential endogeneity biases) to measure the relationship of two variables, our experiment provides causal evidence purely on the role of feedback (absent of monetary confounds) on real-effort task performance. We provide further discussion of the policy-making implications and possible behavioural factors, explaining our findings later in the paper.

Our paper is organised as follows. Section 2 outlines the experimental design and procedures. Section 3 outlines the behavioural hypotheses. Section 4 presents the experimental results. Section 5 discusses the findings, and Section 6 concludes.

2. Experimental design and procedures

2.1. Experimental design

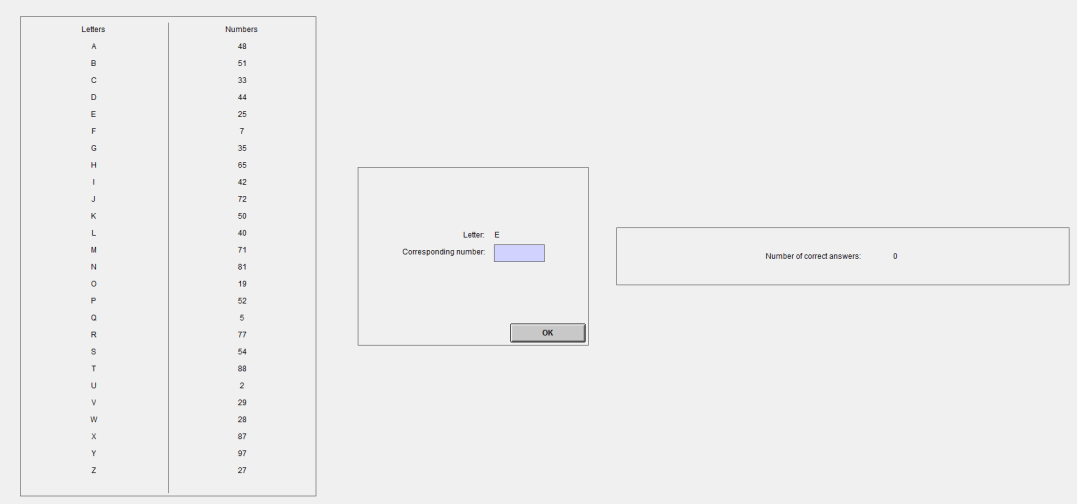
Our experiment consists of three treatments with 10 rounds each and is based on a between-subjects design. In our baseline treatment, subjects are randomly matched into groups of four, and the group composition remains the same throughout the experiment. Subjects have to perform the so-called "encryption task" (previously also used in Erkal et al., 2011; 2018; Benistant & Villeval, 2019) during a maximum period of 90 seconds.[7] The task comprises decoding sets of numbers into letters from a grid of letters that is displayed on a computer screen. If a submitted answer is incorrect, the same letter appears again until the correct answer is provided. The reason for choosing this task is that it is quite boring and intended to induce sufficient disutility. Additionally, no previous knowledge or particular abilities are required to perform the task and learning effects should be minimal. Subjects are paid a flat wage of 10 experimental currency units (ECUs) at the beginning of each round. It is common knowledge that wage is independent of performance. Implementing a flat-wage scheme with equal wages allows us to explore the pure effects of feedback quality on performance, ruling out competitive preferences stemming from monetary motives.

In each round, subjects can solve as many problems as they wish. They can also stop working at any time during the experiment, they can resume work at will, and they can choose

[7] We did not provide subjects with leisure activities, as the work phase was too short for subjects to seriously engage in a leisure task. In addition, recent evidence suggests that the effects of including outside options are less important when subjects perform short real-effort tasks (see Corgnet et al., 2015).

not to perform the task at all. Subjects are continuously informed of their performance, through a score displayed on their screen's interface. Once the 90 seconds elapse, they proceed to the next round in which they have to perform the same task, with the score being reset to zero. Figure 1 provides an example screenshot of the subjects' interface for the encryption task.

Figure 1. Example screenshot for the encryption task



Our treatments differ with respect to the quality of feedback that subjects receive at the end of each round. In the baseline treatment, to which we refer as "No Feedback" (NF, thereafter) treatment, subjects do not receive any feedback about how their performance compares to that of the other three group members.[8] Our experiment consists of two main further treatments where subjects receive feedback about how their rank relates to that of others in their group. In the "High-Quality Feedback" (HQF, thereafter) treatment, subjects are informed of their exact rank relative to the other three group members. In particular, depending on their performance in the task, subjects are told whether they are ranked first (highest performance), second, third, or fourth (lowest performance) in their group.

In the "Low-Quality Feedback" (LQF, thereafter) treatment, subjects are informed of their rank relative to others in their group in a less precise manner. Specifically, subjects are told how well they are (or are not) performing the task within their group. We use the same correspondence between the ranks in the HQF and LQF treatments as shown in Table 1. The key characteristic of the type of feedback in the LQF treatment is that it is difficult for subjects to assess their actual relative rank as is explicitly done in the HQF treatment. In both treatments

[8] We note that having a treatment where subjects are presented with a message unrelated to their performance would lack external validity and thus, a "no feedback" treatment is a more appropriate baseline.

where feedback is provided, subjects are only informed of their relative rank anonymously and cannot see the performance of the other group members.

Table 1. Feedback provision across treatments

Actual rank	No Feedback	High-Quality Feedback	Low-Quality Feedback
First	–	You are ranked 1st in your group.	You are doing very well in your group.
Second	–	You are ranked 2nd in your group.	You are doing well in your group.
Third	–	You are ranked 3rd in your group.	You are not doing well in your group.
Fourth	–	You are ranked 4th in your group.	You are not doing very well in your group.

2.2. Procedures

In total, 248 subjects took part in our experiment. Of these, 48 participated in the NF treatment, 104 participated in the HQF treatment, and 96 in the LQF treatment.[9] All subjects were recruited at the University of Birmingham, using the ORSEE software program (Greiner, 2015) and the experiment was computerised and programmed with the z-Tree software (Fischbacher, 2007). Subjects received the instructions reproduced in Appendix A. After the experiment, subjects were asked to complete a short questionnaire eliciting basic information, such as the subjects' gender, field of study, and nationality. At the end of each session, the subjects were privately paid. The average earnings (including a show-up fee of £2.50) were £10.55 and sessions lasted 50 minutes, on average.[10]

[9] We note that we collected more data for the "High-Quality Feedback" and the "Low-Quality Feedback" treatments relative to the "No Feedback" baseline treatment. This is to ensure we have sufficient power to study how subjects respond to the type of feedback, which is our main concern.

[10] At the end of each round, we also asked subjects to indicate their expected rank in their group. In particular, subjects were asked to indicate a number from "1" (meaning that they are the first in their group) to "4" (meaning that they are the last in their group). If their estimate was correct, they received an extra £0.10.

To formally test whether the subjects' gender, field of study, and nationality are balanced across treatments, we conduct balance checks, the results of which are reported in Table 2. The average values of each demographic characteristic, along with the corresponding standard deviations for each treatment, are shown in columns 1–3. We perform two-sided Kruskal–Wallis tests to check whether the distribution of given demographic variables differs across treatments, and we do not find evidence that this is the case ($p > 0.432$). When we perform two-sided Mann–Whitney tests, checking for pairwise treatment differences, we reach the same conclusions, indicating that subjects' observable characteristics were similar across treatments.[11]

Table 2. Balance checks of individuals' characteristics across treatments

	NF treatment	LQF treatment	HQF treatment	Kruskal–Wallis test	Mann–Whitney test NF vs. LQF	Mann–Whitney test NF vs. HQF	Mann–Whitney test HQF vs. LQF
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Female	0.50	0.58	0.58	0.689	0.345	0.377	0.927
	(0.50)	(0.49)	(0.49)				
Economics degree	0.21	0.33	0.26	0.432	0.121	0.495	0.255
	(0.41)	(0.47)	(0.44)				
Nationality	0.63	0.66	0.69	0.785	0.713	0.413	0.587
	(0.48)	(0.48)	(0.46)				
Obs.	48	96	104				

Notes: The table reports information on the students' demographic variables such as gender, field of study, and nationality. "Female" is a dummy variable that takes the value 1 if a subject is female and 0 otherwise. "Economics degree" is a dummy variable which takes the value 1 if a subject is studying for an Economics/Business degree and 0 otherwise, and "Nationality" takes the value 1 if a subject comes from the UK or another European country and 0 otherwise. Columns 1–3 report average and standard deviations (in parentheses) of each demographic variable across treatments. Column (4) reports two-sided p-values from a Kruskal–Wallis test. Columns 5–7 report two-sided p-values from a Mann–Whitney test.

[11] In Appendix B, we also provide the correlation matrices of the variables reported in the paper (see Table B.1).

3. Behavioural hypotheses

Under the assumption that individuals maximise their own payoff, the theoretical prediction in all treatments is that the minimum effort should be exerted. However, as past research has shown (e.g., Charness et al., 2014; Kuhnen & Tymula, 2014), participants may have an intrinsic motivation for working. Evidence from gift exchange game experiments shows that workers provide positive effort levels even if wages are flat (e.g., Cohn et al., 2015; Gächter & Thöni, 2010; Kube et al., 2013). Relying on this literature, we formulate the following hypothesis:

Hypothesis 1: Even in a flat-wage environment, individuals exert positive levels of effort.

The provision of rank feedback can strengthen intrinsic motivation concerns that can be due to social comparisons, desire for dominance, and acquiring status within their group. In fact, there is expansive experimental research offering evidence that individuals care not only about their own payoffs but also about social image and status (e.g., Ball & Eckel, 1998; Rustichini, 2008; Eckel et al., 2010; Kosfeld & Neckermann, 2011). Kuhnen and Tymula (2012) show that agents work harder when they observe their ranking, and highlight the key role of self-esteem motives in driving performance differences. In a flat-wage scheme, Falk and Ichino (2006) and Mas and Moretti (2009) underline the role of peer effects in raising productivity. Based on previous findings, we can write the following hypothesis:

Hypothesis 2: In a flat-wage environment, providing rank feedback increases performance.

However, the existing literature does not take into account the role that the quality of feedback may play in determining performance levels. Our last hypothesis concerns the effects of feedback quality on performance and corresponds to the novel hypothesis being tested in this experiment. To the best of our knowledge, we are not aware of any studies on the impact of feedback quality on individuals' performance. Motivated by Gibbs' work (1991), suggesting that "those with the greatest reason to run as fast as they can are those who are in a tight race for the finish", we assume that the provision of vague rank feedback crowds out individuals' intrinsic motivation. For example, an individual may be reluctant to work when the rank feedback received is not precise because this may weaken their concerns for social image and/or status in the group. Alternatively, self-esteem concerns and desire to outperform others may be less salient when individuals do not know exactly how they rank in relation to their peers.

Our third hypothesis is organised by the feedback intervention theory put forward by Kluger and DeNisi (1996). In particular, this framework stresses the key role that feedback plays in affecting performance and posits that feedback cues may improve or debilitate performance, depending on whether feedback cues direct attention to task-learning/task-motivation processes or meta-task processes, respectively.[12] When attention is shifted to task-relevant processes, performance increases, as cognitive resources are not depleted to processes away from the details of the task. We expect that high-quality feedback enhances task involvement and subsequently performance. In contrast, the provision of low-quality feedback is expected to direct attention to meta-task goals and lead to disengagement from the task, bringing about negative performance effects. Based on the conceptual framework of Kluger and DeNisi's (1996) feedback intervention theory, we formulate Hypothesis 3:

Hypothesis 3: In a flat-wage environment, providing high-quality rank feedback increases performance in relation to the provision of low-quality feedback.

4. Experiment results

In discussing our results, we start by examining whether average performance differs across our treatments and, if so, how, testing for our behavioural hypotheses. To gain a deeper understanding of the effects of feedback quality on performance, we also explore whether feedback quality affects the percentage of those who put in zero effort across treatments. Finally, we test whether semantics in the low-quality feedback matter. The analysis presented in this section has been done using Stata (version 15).

4.1. Performance across treatments

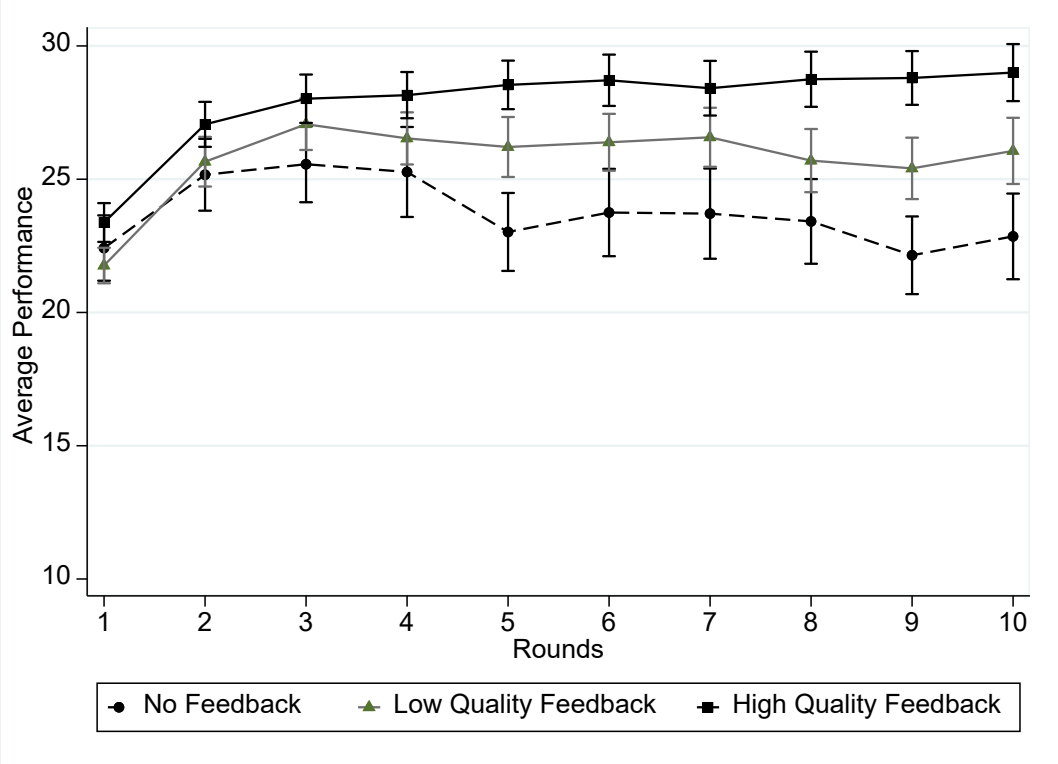
Figure 2 presents the average performance of subjects in each round across treatments (including 95% confidence intervals). In all three treatments, we observe a jump in performance between the initial and second rounds, probably due to a learning effect.[13] This is

[12] Following Kluger and DeNisi's (1996) definition, meta-task processes include "processes that link the focal task with higher order goals, such as the evaluation of the implication of task performance for the self".

[13] Looking only at the first round, we find that average performance is 22.42 correct answers in the NF treatment, 23.38 correct answers in the HQF treatment, and 21.76 correct answers in the LQF treatment. A Kruskal–Wallis test confirms that there are no differences in the initial performance across treatments ($p = 0.236$). All the tests reported in this section correspond to two-sided tests. Each matching group is treated as the independent unit of observation. The number of matching groups (clusters) is equal to 12 in the NF treatment, 24 in the LQF treatment, and 26 in the HQF treatment.

in line with previous experimental studies that have used the same task (see Erkal et al., 2011; Charness et al., 2014). Over time, average performance slightly declines for the LQF treatment, but this effect is stronger for the NF treatment. By contrast, we observe that in the HQF treatment, average performance slightly increases as the session progresses.[14] Looking at the aggregate data across all 10 rounds, our analysis shows that, on average, subjects provide 23.73 correct answers (s.d.=4.77) in the NF treatment, 25.73 correct answers (s.d.=3.51) in the LQF treatment, and 27.88 correct answers (s.d.=2.97) in the HQF treatment. Our analysis shows that even in a flat-wage environment, subjects are willing to exert positive effort in all treatments. This provides support in favour of Hypothesis 1.

Figure 2. Average performance over time across treatments



By performing a Mann–Whitney test, we find that subjects' performance is significantly higher in the HQF treatment compared to the NF treatment ($p = 0.012$) and the LQF treatment ($p = 0.034$). When we compare average performance between the LQF

[14] We test more formally for linear trends in each treatment by performing three separate OLS regressions for each treatment (see Table B.2 in Appendix B). In each of these regressions, the dependent variable is the number of correct answers provided by a subject and, as an independent variable, we include a "Round" variable. We find that in the NF treatment, the coefficient of the "Round" variable is not statistically significant, whereas in the HQF treatment it is positive and statistically significant (at the 1% level). For the LQF treatment, we find a marginally (significant at the 10% level) positive trend of performance over time. Results remain the same when we replace the "linear round trend" variable with period dummies (see Table B.3 in Appendix B).

treatment and the NF treatment, we find no statistically significant difference at conventional levels ($p = 0.180$). These results indicate that receiving high-quality feedback has significant positive effects by increasing performance. Additionally, we fail to reject the null hypothesis that subjects perform equally when they receive low-quality feedback versus no feedback at all. Our results show that what mainly matters for increasing subjects' performance is the quality of the feedback provided. Unlike the high-quality rank feedback which significantly enhances performance, the provision of low-quality rank feedback increases performance levels but not significantly so when compared with the no-feedback treatment – a finding that provides only partial support for our Hypothesis 2. In contrast, performance is significantly higher when high-quality feedback is provided compared to either of the other two treatments, which reinforces our Hypothesis 3.

Next, we look separately at subjects' performance in the first five and last five rounds across treatments. Starting with the first five rounds, we observe that subjects provide, on average, 27.03 correct answers (s.d.=2.74) in the HQF treatment, 25.44 correct answers (s.d.=3.37) in the LQF treatment, and 24.29 correct answers (s.d.=4.39) in the NF treatment. A Mann–Whitney test shows that subjects perform better in the HQF treatment when compared to the NF treatment, even though this difference is weakly significant ($p = 0.064$). By contrast, we do not find any statistically significant differences in the average performance of the first five rounds between the NF and LQF treatments ($p = 0.524$). When comparing average performance between the LQF and the HQF treatments, we only observe marginally insignificant differences ($p = 0.125$). Turning to performance in the last five rounds, we observe that subjects provide, on average, 28.73 correct answers (s.d.=3.64) in the HQF treatment, 26.03 correct answers (s.d.=4.08) in the LQF treatment, and 23.18 correct answers (s.d.=5.48) in the NF treatment. By performing a Mann–Whitney test, we observe statistically significant differences between the HQF treatment and both the NF treatment ($p = 0.004$) and the LQF treatment ($p = 0.024$) and a weakly significant difference between the NF and the LQF treatments ($p = 0.097$).

It is also interesting to examine average performance in the last round as this would provide evidence of whether endgame effects are observed across treatments. Specifically, the average final performance is 29 correct answers (s.d.=3.79) in the HQF treatment, 26.06 correct answers (s.d.=4.79) in the LQF treatment, and 22.85 correct answers (s.d.=6.61) in the NF treatment. A Mann–Whitney test reveals that there still exist significant differences in average performance between the HQF treatment and the NF treatment ($p = 0.009$) and between the

HQF and the LQF treatments ($p = 0.027$). By contrast, we find that the difference in average performance between the LQF and the NF treatments is not statistically significant ($p = 0.118$). Overall, this analysis indicates that even in the very final round, the provision of high-quality feedback still has significant positive effects on average performance compared to either of the other two treatments.

The results of the non-parametric tests are corroborated by OLS regression analyses, reported in Table 3, where we regress the number of correct answers on the treatment dummies (the baseline being the NF treatment) and the number of rounds (capturing potential time trend effects). We perform four separate regressions, in which we look at treatment differences in performance across all rounds, rounds 1–5, rounds 6–10, and rounds 2–10. We find that the HQF treatment yields higher effort levels than both the NF and the LQF treatments, whereas there are no statistically significant differences in performance between the NF and the LQF treatments. We also find that the variable "Rounds" is positive and statistically significant only when we consider all rounds and the first five rounds, but not for the last five rounds or rounds 2–10. This implies that the initial period effects in average performance disappear as the session progresses, and subjects' average performance remains stable for most of the time throughout the session. We also report OLS regression analyses (see Table B.4 in Appendix B), where we control for subjects' demographic variables. We obtain similar results when we replace the variable "Rounds" with separate period dummies (see Table B.5 in Appendix B), and we also control for demographics (see Table B.6 in Appendix B). Our treatment effects remain robust when we include a quadratic trend (see Table B.7 in Appendix B) and further control for demographic variables (see Table B.8 in Appendix B).

Table 3. Performance differences by treatment

Dependent variable: Number of correct answers				
Independent variables	All Rounds	Rounds 1–5	Rounds 6–10	Rounds 2–10
LQF	2.003	1.156	2.850	2.299
	(1.506)	(1.402)	(1.737)	(1.576)
HQF	4.150**	2.741*	5.560**	4.505**
	(1.449)	(1.336)	(1.685)	(1.520)
Rounds	0.205**	0.883***	-0.095	-0.025
	(0.068)	(0.130)	(0.100)	(0.068)
Constant	22.602***	21.640***	23.933***	24.024***
	(1.280)	(1.168)	(1.582)	(1.356)
Observations	2,480	1,240	1,240	2,232
Differences in average performance between treatments				
LQF minus HQF	-2.147*	-1.585†	-2.710*	-2.207*
	(0.913)	(0.864)	(1.085)	(0.945)

Notes. OLS regressions. Robust standard errors, clustered at the matching group level (62 clusters), are reported in parentheses.

† $p < .10$

* $p < .05$

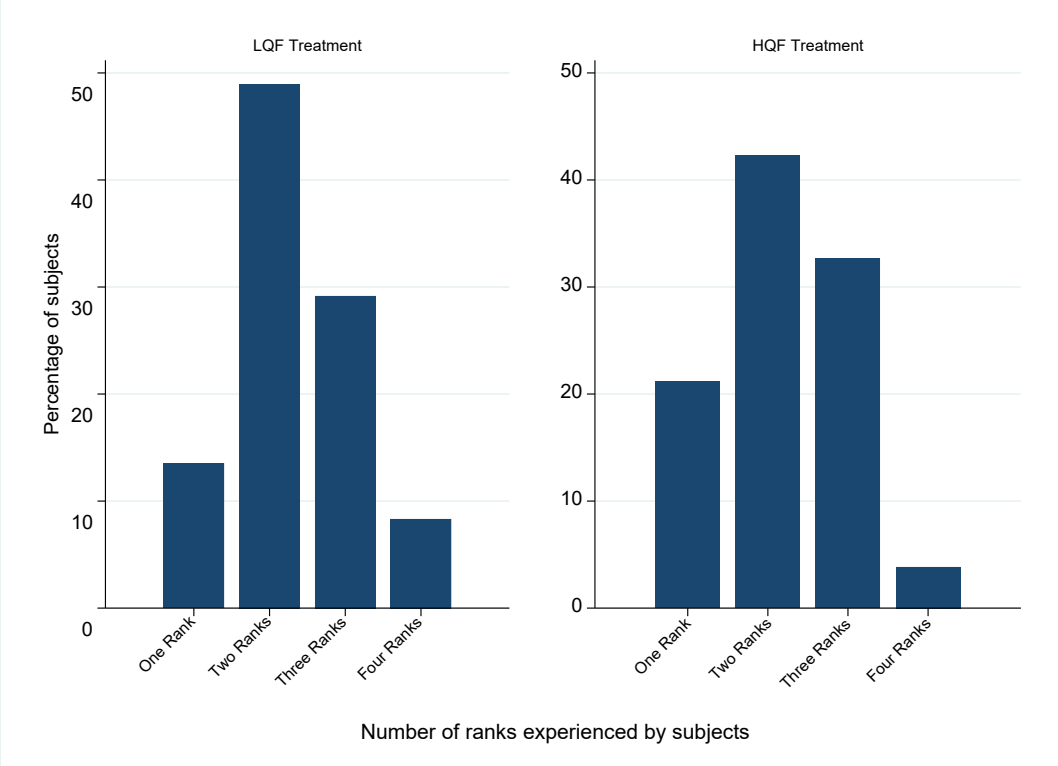
** $p < .01$

*** $p < .001$

One potential reason why we observe such performance differences between the HQF and the LQF treatments may be that subjects in the latter treatment may attempt to experience different ranks in order to identify which rank corresponds to which type of feedback (given that in the LQF treatment there is an unclear correspondence between the two). To test for this, we look at the number of ranks that subjects experience in each treatment and provide evidence that the number of ranks that subjects see is the same between the HQF and the LQF treatments. Figure 3 shows the distribution of the number of ranks that subjects see in the HQF and the LQF treatments throughout all rounds. In both treatments, the majority of subjects experience two ranks (42.31% and 48.96% for the HQF and LQF treatments, respectively), whereas the percentage of subjects experiencing all four ranks is the lowest (3.85% and 8.33% for the HQF

and LQF treatments, respectively).[15] We find that both distributions are not significantly different from each other (p = 0.380), implying that in the LQF treatment the number of ranks experienced by subjects is similar in comparison to the HQF treatment. Our analysis indicates that the treatment differences observed in average performance between the HQF and LQF treatments are not due to subjects trying to find out which rank corresponds to which type of feedback.

Figure 3. Distribution of the number of ranks experienced by subjects in the HQF and LQF treatments



4.2. Provision of zero effort across treatments

In this section, we explore the effects of feedback quality on subjects putting in zero effort. Specifically, we define "zero effort" by taking into account those subjects who have provided both zero correct and zero wrong answers in the encryption task. Table 4 shows the average percentage of subjects who put zero effort into each round separately across treatments. We observe that the percentage of subjects who make no effort in the HQF treatment is relatively

[15] For completeness, we include the corresponding figure for the NF treatment in Appendix B (see Figure B.1). In this treatment, where subjects do not receive any feedback about their rank, we observe that the modal number of ranks that subjects experience is either one or two (approximately 35% in each case).

low, starting at 1.92% in round 1 and rising to only 3.85% in round 10. Overall, in the HQF treatment, the percentage of subjects putting in zero effort remains very low and stable across rounds. However, this is not the case in either of the other two treatments. In the LQF and the NF treatments, we find that on average 2.08% of subjects expend no effort at all in round 1. However, in the last round, this percentage increases more than five times: 10.42% of subjects in the LQF treatment and 12.5% of subjects in the NF treatment put in zero effort.

Table 4. Percentage of subjects who put in zero effort across treatments

	Rounds										
	1	2	3	4	5	6	7	8	9	10	Total
NF	2.08	4.17	6.25	6.25	10.42	10.42	10.42	8.33	14.58	12.50	8.54
LQF	2.08	2.08	4.17	1.04	7.29	4.17	5.21	10.42	10.42	10.42	5.73
HQF	1.92	1.92	1.92	0.96	1.92	1.92	2.88	3.85	2.88	3.85	2.40

Across all rounds, the percentage of subjects putting in no effort is 8.54% in the NF treatment and 5.73% in the LQF treatment. A Mann–Whitney test indicates no significant differences between these two treatments ($p = 0.471$). However, when we compare the percentage of subjects putting in zero effort in the HQF treatment and either of the other two treatments, we find statistically significant differences. In particular, the percentage of subjects expending no effort in the HQF treatment is significantly lower with regard to the NF treatment ($p = 0.024$), as well as the LQF treatment ($p = 0.024$). Our results are corroborated by a Probit regression analysis reported in Table B.9 in Appendix B. Overall, this analysis finds evidence that the provision of high-quality feedback has a significantly positive effect on reducing the number of subjects who slack off by putting in no effort at all when performing a task. This also highlights the importance of high-quality feedback in reducing zero-effort choices in a flat-wage environment where free-riding incentives are dominant.

4.3. The role of semantics

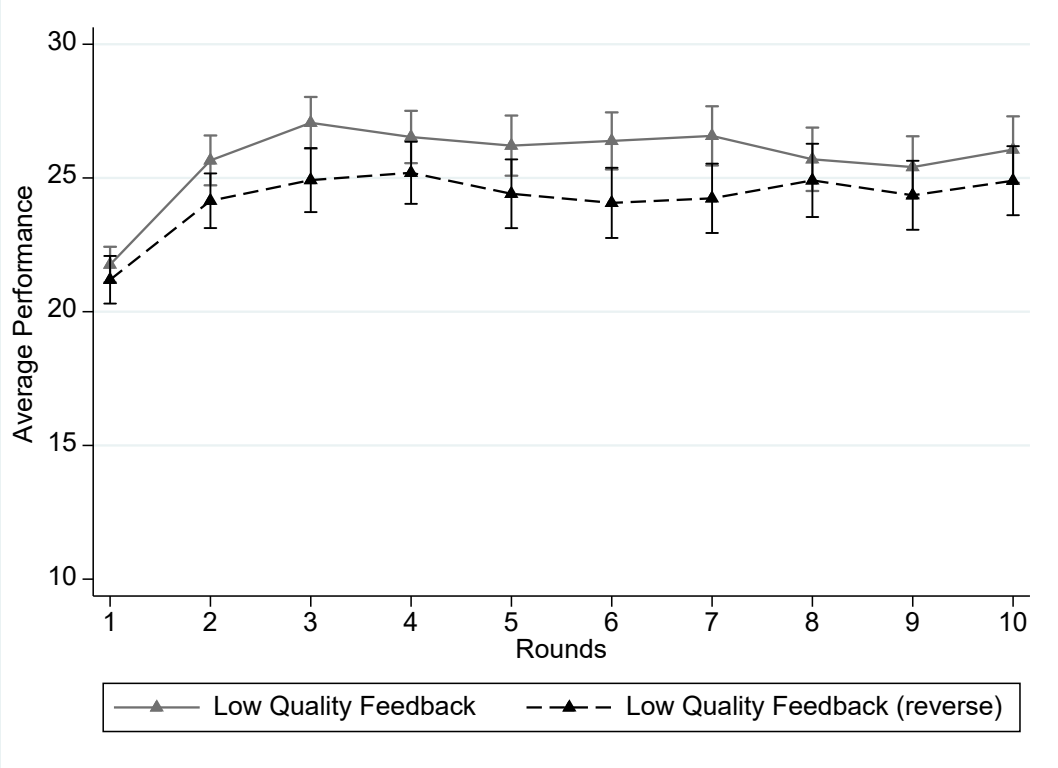
In a separate series of experiments, we run new sessions aimed at exploring the role of semantics in the LQF treatment which may play a role in determining subjects' performance. For example, it might be the case that when subjects are provided with low-quality feedback, the wording used for the last two ranks (i.e., "You are not doing well in your group" and "You are not doing very well in your group") may not unanimously correspond to the third and fourth ranks, respectively. Instead, subjects could have perceived that these two ranks are in reverse

order. To test whether the semantics of the order of the last two ranks affects performance in the LQF treatment, we conducted additional experiments whereby the statements "You are not doing very well in your group" and "You are not doing well in your group" are in reverse order and correspond to the third and fourth ranks, respectively. All other aspects of the new experiment are identical to the LQF treatment as presented in Section 2. By comparing subjects' performance in the LQF treatment when these two separate orders for the last two ranks have been used, we can measure the potential scope of semantics in changing behaviour. We refer to the LQF treatment that uses the new reversed order of the last two ranks as the "LQF (reverse)" treatment. Table B.10 in Appendix B shows that subjects' characteristics in the LQF (reverse) treatment and the LQF treatment are not statistically distinguishable.

Figure 4 illustrates the average performance of subjects across rounds between the two LQF treatments (including 95% confidence intervals). Similarly to the LQF treatment, we observe a jump in performance between the initial and second rounds.[16] In addition, looking at the aggregate data across all 10 rounds, our analysis shows that, on average, subjects provide 24.23 correct answers (s.d.=5.77) in the LQF (reverse) treatment as compared to 25.73 correct answers (s.d.=3.51) in the LQF treatment. A Mann–Whitney test confirms that subjects' performance is not significantly different at conventional levels between the two LQF treatments ($p = 0.482$).

[16] Looking only at the first round, we find that the average performance is 21.76 correct answers in the LQF treatment and 21.19 in the LQF (reverse) treatment. A Mann–Whitney test confirms that there are no statistically significant differences in the initial performance between the two treatments ($p = 0.886$). All tests reported correspond to two-sided tests. Each matching group is treated as the independent unit of observation.

Figure 4. Average performance over time for the two LQF treatments



Additionally, we find that performance is not statistically distinguishable when we compare average correct answers in the first five and last five rounds across treatments. Specifically, when we look at the first five rounds, subjects, on average, provide 23.97 correct answers (s.d.=5.22) in the LQF (reverse) treatment as compared to 25.44 correct answers (s.d.=3.37) in the LQF treatment ($p = 0.410$). Turning to the last five rounds, we find that in the LQF (reverse) treatment, subjects, on average, provide 24.49 correct answers (s.d.=6.59) as compared to 26.03 correct answers (s.d.=4.08) in the LQF treatment ($p = 0.475$).

We also perform OLS regressions where the dependent variable corresponds to a subject's number of correct answers. As an independent variable, we include a dummy variable that is equal to 1 for the LQF (reverse) treatment (with the baseline category being the LQF treatment). We perform four separate regressions, in which we look at treatment differences in performance across all rounds, rounds 1–5, rounds 6–10, and rounds 2–10. Table 5 reports the results of our regressions.

Table 5. Performance differences by LQF treatments

Dependent variable: Number of correct answers				
Independent variables	All Rounds	Rounds 1–5	Rounds 6–10	Rounds 2–10
LQF (reverse)	-1.501	-1.471	-1.532	-1.605
	(1.407)	(1.295)	(1.617)	(1.464)
Rounds	0.188*	0.867***	-0.010	-0.036
	(0.073)	(0.129)	(0.109)	(0.075)
Constant	24.702***	22.842***	26.103***	26.392***
	(0.737)	(0.745)	(1.109)	(0.782)
Observations	1,840	920	920	1,656

Notes. OLS regressions. Robust standard errors, clustered at the matching group level (62 clusters), are reported in parentheses.

† $p < .10$

* $p < .05$

** $p < .01$

*** $p < .001$

Our regression analyses corroborate our findings from our non-parametric analysis. In particular, we find statistically insignificant differences across treatments in each of our four regression models. We also observe that subjects' performance increases in the first five rounds but remains stable in the second half of the session, or even when we consider rounds 2–10 (similar to the period effects reported in Table 3). Taken together, our analysis indicates that semantics do not play a significant role in affecting performance in the LQF treatment, as the order of ranks in which the last two feedback statements correspond does not affect behaviour in a statistically significant way.

5. Discussion and policy implications

In this section, we provide some discussion of the main mechanisms that may explain our results. The main finding of our study highlights the key role of feedback quality in determining workers' performance. We show that it is not only the provision of feedback but, importantly, it is the quality of feedback that makes workers more productive: high-quality feedback raises

performance compared to both the provision of low-quality or no feedback. However, what might explain the effectiveness of high-quality feedback?

We first discuss the mechanisms that our experimental design rules out. In our setting, individuals' relative performance/rank is not exposed to an audience, allowing us to rule out social image motives as important drivers of behaviour (see Ariely et al., 2009). Additionally, the influence of monitoring effects is also excluded (e.g., Dickinson & Villeval, 2008), as the provision of feedback was private and anonymous and cannot be observed by other group members.[17] Furthermore, as the task that subjects work on does not involve changes in strategy and requires no particular abilities, we exclude confounds that may come from learning effects due to the nature of the task. Feedback may also influence output if compensation is tied to performance. There is expansive empirical evidence that external interventions (output-based, for example) crowd out intrinsic motivation and undermine productivity (e.g., Frey, 1997; Deci et al., 1999; Frey & Jegen, 2001). Bénabou and Tirole (2003) formalise the concepts of intrinsic and extrinsic motivation and show under which conditions the latter will "crowd out" or "crowd in" the former. Our design rules out the role of extrinsic rewards interacting with the provision of feedback, indicating that the quality of (non-financial) incentives crowds out intrinsic motivation.

Next, we turn to possible mechanisms that can explain the observed patterns of behaviour. In our setting, subjects are told their own relative rank in their groups and, thus, individuals may care about their relative status. It has been argued in the literature that individuals may derive utility from relative status per se (e.g., Frank, 1985; Robson, 1992). Our results show that subjects in the LQF (and NF) treatment perform significantly worse than those in the HQF treatment, implying that status-seeking individuals are more motivated to work hard in the latter treatment where relative rank can be easily identified (HQF treatment) as opposed to the treatments where this is harder to identify (NF and LQF treatment). Thus, in the latter treatments, subjects would show less regard for relative status due to the lack of a clear one-to-one correspondence between performance and rank. Since the subjects' performances are not tied to their earnings, our findings also show that relative status concerns may not necessarily be driven by monetary incentives but could solely be due to intrinsic motives. Our results are in line with previous work showing the positive effects of relative rank

[17] The influence of peer effects (e.g., Falk & Ichino, 2006; Mas & Moretti, 2009) is also likely to be minimal but cannot be fully excluded, as when subjects perform the task, they belong to groups.

feedback on performance. For example, as Kuhnen and Tymula (2012) propose, offering feedback may modify individuals' self-esteem. Consonant with their findings, we observe that individuals work harder when they learn their ranking (in particular, when high-quality feedback is provided), relative to when no feedback is provided. This finding may be attributed to self-esteem considerations that might be more salient when high-quality feedback is given. Specifically, providing subjects with their own relative rank may capture the fact that self-esteem can be determined by relative comparisons (e.g., Szymanski & Harkins, 1987). In contrast, self-esteem motives can be nullified when the quality of feedback is low, in which relative rank comparisons cannot be easily interpreted and, thus, relative self-evaluation comparisons cannot be easily made. Alternatively, status-seeking individuals may also be simply motivated by a desire to dominate in competitive environments (as discussed by Charness & Grosskopf, 2001; Charness & Rabin, 2002). These motives may be triggered more prominently when high-quality feedback is provided, as knowing their exact standing in their group may enhance competition for status, leading to higher performance.

Although we do not directly manipulate self-image concerns (as in Falk, 2020), status-seeking may modify self-awareness (i.e., how one is feeling about oneself) by varying feedback quality. Self-image considerations have mainly been the focus of psychology research, with existing evidence showing that greater self-awareness has positive effects on behaviour, such as less cheating (see Vallacher & Solodky, 1979; Mazar et al., 2008). In our set-up, providing information about their exact relative rank in the group may raise individuals' awareness of their own choices, resulting in significant performance differences across treatments.

Moreover, performance appraisal is a highly emotional process, and the presence of vague performance reviews may further cause negative emotional reactions, such as aggravation and anger. The trigger of negative emotions due to the lack of (clear) feedback may be a likely mechanism explaining why, in the HQF treatment, individuals outperform those in the LQF or the NF treatment. Recent experimental research provides evidence that induced anger negatively affects productivity in the presence of extrinsic motives (see Oswald et al., 2015). However, further research is warranted to examine more formally the link between negative emotions and productivity when confounds from financial incentives have been removed. In a similar vein, Azmat and Iriberri (2016) provide evidence that the relative performance feedback effects may be emotions-based. Relatedly, using fMRI techniques, recent research shows that the joy of outperforming others is linked with the activation of the neural circuitry associated with reward processing (Dohmen et al., 2011). This points to a

direction in which the differential effects of feedback quality on performance may potentially interact with emotions.

Before we move to a summary of the concluding remarks for our study, we reflect on the applied implications of our feedback intervention. Organisations must know whether and how they can improve their total output through the provision of feedback. The key finding of our experiment is that high-quality feedback boosts performance compared to low-quality feedback when monetary concerns are ruled out. In light of this finding, our experiment has implications for the characteristics of optimal corporate feedback policies. Our results demonstrate that an intervention that contains feedback cues supporting learning and increasing motivation for task-relevant aspects – as our high-quality feedback treatment does – can yield significant performance gains. Yet, if the feedback intervention affects performance through meta-task processes (as in our low-quality feedback treatment), workers' performance will be debilitated. This has implications for the design of work environments, which should implement high-quality feedback interventions to encourage task learning and motivations that will, in turn, have positive effects on performance.

The use of incentive plans adopted by modern organisations to boost employees' performance and career prospects is a common management practice and, typically, incentive schemes rely on the use of extrinsic rewards (see, for example, Lazear (2000) on piece-rates and Harbring & Irlenbusch (2003) on tournaments). While tying performance to pay might have positive effects in raising productivity, existing evidence demonstrates that financial incentives are not always an ideal motivator and can sometimes backfire (e.g., Deci, 1971; Ariely et al., 2009). In addition, financial incentive schemes may require costly institutional changes that may be difficult to implement on practical grounds. Our paper uncovers the causal impact that a cost-effective mechanism has on employees' productivity, highlighting the role of non-financial incentives that can be used by firms to achieve better outcomes. Our focus is on feedback quality. We find that such an inexpensive tool can generate impressive performance gains. Specifically, leaders and work managers who adopt a cost-effective feedback intervention promoting rank precision will, thus, document impressive performance gains. It is therefore important that in organisational settings, where the provision of feedback constitutes an important aspect in determining employees' future career prospects and salary negotiations (such as in performance appraisals), team leaders and managers offer specific and high-quality feedback that can spur employees' motivation.

Even though our experimental set-up does not allow us to capture the complex and multifaceted aspects of feedback policies, our results have clear implications for enhancing productivity in organisational settings. Actions that are undertaken towards improving performance appraisals processes, through clarity and transparency of feedback policies, are more likely to yield significant performance gains for firms.

6. Conclusions

Performance appraisals constitute a common practice in organisational life and have been traditionally used to promote employee performance in modern corporations. In this paper, we report on an experiment designed to directly test for the causal impact that feedback quality has on individuals' productivity in a flat-wage environment (ruling out financial considerations from the decision-making setting). This allows us to focus on whether the quality of non-financial incentives crowds in or crowds out workers' output. In our two main treatments, feedback quality varies in that subjects know ("High-Quality Feedback") or do not know ("Low-Quality Feedback") their exact relative rank in their group. Our main findings indicate the causal impact of feedback on task performance: providing workers with low-quality feedback decreases their intrinsic motivation and leads to significantly less output as compared to a high-quality feedback setting. In addition, we note that the type of feedback we implemented in our experiment may also be applicable to other contexts such as in sports and school grades. Our experiment demonstrates the positive effects that high-quality feedback has on productivity, implying that in the presence of free-riding incentives, firms should make efforts to improve the feedback quality offered to employees as this can be a tool that can motivate individuals to work harder.

Our study differentiates us from the expansive literature in behavioural management and applied psychology looking at the interplay between feedback intervention effects and performance and establishes a causal link between the two. Past management studies rely on the use of questionnaire techniques that do not allow researchers to cleanly identify causality because the measured variables may be endogenous (that is, they may depend on and correlate with other variables). Having endogenous variables as independent regressors generates inconsistent estimates and does not capture the true causal effect (see Antonakis et al., 2014; 2016). In our experiment, we exogenously manipulate the provision of feedback across treatments, rendering feedback – our main variable of interest – an exogenous variable. This

enables us to causally attribute the observed effects on task performance behaviour to the role of feedback provision, varying exogenously across treatments. This allows us to discuss the policy implications of our findings, as we can establish the direction of the relationship between feedback provision and task performance.

Like all studies in social sciences, this study is not without limitations. Feedback quality may be perceived differently by subjects (see Nae et al., 2015) and, thus, future research would be interesting to assess whether subjects perceive high (low) feedback quality as being more (less) useful in driving their performance accordingly. For instance, our findings may be driven by the fact that one type of feedback is more detailed than the other; alternatively, they might also be driven by the fact that the high-quality feedback is perceived as an objective piece of information, whereas the low-quality feedback is perceived as a message with either negative or positive valence which either demotivates or encourages the subjects to put in more effort. Furthermore, we note that verbal feedback varies in several ways with respect to numerical feedback (e.g., mix of negative and positive framing and feedback "precision" or "objectivity"). Which aspects of low-quality feedback drive performance differences warrants further research.

Nevertheless, we see our findings as a potentially important tool at the hands of leaders and human resource managers. Overall, our study highlights the crucial role of providing high-quality feedback as a mechanism that can be used by firms to encourage higher levels of performance. On the contrary, when organisations provide poor feedback, workers are being demotivated from generating higher outputs. Our findings provide modern organisations that employ performance appraisals as part of evaluating their employees with a useful cost-effective policy instrument to boost their productivity levels.

References

Antonakis, J., Bastardoz, N., Jacquart, P., & Shamir, B. (2016). Charisma: An ill-defined and ill-measured gift. *Annual Review of Organizational Psychology and Organizational Behavior*, 3, 293–319.

Antonakis, J., Bendahan, S., Jacquart, P., & Lalive, R. (2014). Causality and endogeneity: Problems and solutions. In D. V. Day (Ed.), *The Oxford Handbook of Leadership and Organizations* (pp. 93–117). New York: Oxford University Press.

Ariely, D., Bracha, A., & Meier, S. (2009). Doing good or doing well? Image motivation and monetary incentives in behaving prosocially. *American Economic Review,* 99(1), 544–555.

Avolio, B. J., Bass, B. M., & Jung, D. I. (1995). *MLQ Multifactor Leadership Questionnaire: Technical Report*. Redwood City, CA: Mindgarden.

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.

Azmat, G., & Iriberri, N. (2016). The provision of relative performance feedback: An analysis of performance and satisfaction. *Journal of Economics & Management Strategy*, 25(1), 77–110.

Ball, S. B., & Eckel, C. C. (1998). The economic value of status. *Journal of Socio-Economics.* 27(4), 495–514.

Bandura, A. (1991). Social cognitive theory of self-regulation. *Organizational Behavior and Human Decision Processes,* 50(2), 248–287.

Barankay, I. (2012). Rank incentives. Evidence from a randomized workplace experiment. Working paper, Wharton School, University of Pennsylvania.

Bartling, B., & von Siemens, F. A. (2010). The intensity of incentives in firms and markets: Moral hazard with envious agents. *Labour Economics*, 17(3), 598–607.

Bass, B. M., & Avolio, B. J. (1997). *Full Range Leadership Development: Manual for the Multifactor Leadership Questionnaire*. Palo Alto, CA: Mindgarden.

Bénabou, R., & Tirole, J. (2003). Intrinsic and extrinsic motivation. *The Review of Economic Studies*, 70(3), 489–520.

Benistant, J., & Villeval, M. C. (2019). Unethical behavior and group identity in contests. *Journal of Economic Psychology,* 72, 128–155.

Blanes i Vidal, J., & Nossol, M. (2011). Tournaments without prizes: Evidence from personnel records. *Management Science*, 57(10), 1721–1736.

Butler, R. (1987). Task-involving and ego-involving properties of evaluation: Effects of different feedback conditions on motivational perceptions, interest, and performance. *Journal of Educational Psychology*, 79(4), 474–482.

Charness, G., & Grosskopf, B. (2001). Relative payoffs and happiness. *Journal of Economic Behavior and Organization*, 45(3), 301–328.

Charness, G., Masclet, D., & Villeval, M. C. (2014). The dark side of competition for status. *Management Science*, 60(1), 38–55.

Charness, G., & Rabin, M. (2002). Understanding social preferences with simple tests. *Quarterly Journal of Economics*, 117(3), 817–869.

Cohn, A., Fehr, E., & Goette, L. (2015). Fair wages and effort provision: Combining evidence from a choice experiment and a field experiment. *Management Science*, 61(8), 1777–1794.

Corgnet, B., Hernán-González, R., & Schniter, E. (2015). Why real leisure really matters: Incentive effects on real effort in the laboratory. *Experimental Economics*, 18(2), 284–301.

Cropanzano, R., & Mitchell, M. (2005). Social exchange theory: An interdisciplinary review. *Journal of Management*, 31(6), 874–900.

Deci, E. L. (1971). Effects of externally mediated rewards on intrinsic motivation. *Journal of Personality and Social Psychology*, 18(1), 105–115.

Deci, E. L., Koestelr, R., & Ryan, R. M. (1999). A meta-analytic review of experiments examining the effects of extrinsic rewards on intrinsic motivation. *Psychological Bulletin*, 125(6), 627–668.

Dickinson, D., & Villeval, M. C. (2008). Does monitoring decrease work effort?: The complementarity between agency and crowding-out theories. *Games and Economic Behavior*, 63(1), 56–76.

Dohmen, T., Falk, A., Fliessbach, K., Sunde, U., & Weber, B. (2011). Relative versus absolute income, joy of winning, and gender: Brain imaging evidence. *Journal of Public Economics*, 95(3–4), 279–285.

Eckel, C., Fatas, E., & Wilson, R. (2010). Cooperation and status in organizations. *Journal of Public Economic Theory*, 12(4), 737–762.

Erkal, N., Gangadharan, L., & Koh, B. H. (2018). Monetary and non-monetary incentives in real-effort tournaments. *European Economic Review*, 101, 528–545.

Erkal, N., Gangadharan, L., & Nikiforakis, N. (2011). Relative earnings and giving in a real-effort experiment. *American Economic Review*, 101, 3330–3348.

Eriksson, T., Poulsen, A., & Villeval, M. C. (2009). Feedback and incentives: Experimental evidence. *Labour Economics*, 16(6), 679–688.

Eriksson, T., & Villeval, M. C. (2012). Respect and relational contracts. *Journal of Economic Behavior and Organization*, 81(1), 286–298.

Etzioni, A. (1971). *Modern Organizations*. Englewood Cliffs, NJ: Prentice-Hall.

Falk, A. (2020). Facing yourself – A note on self-image. *Journal of Economic Behavior & Organization*, 18(6), 490–516.

Falk, A., & Ichino, A. (2006). Clean evidence on peer effects. *Journal of Labor Economics*, 24(1), 39–57.

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

Frank, R. H. (1985). The demand for unobservable and other nonpositional goods. *American Economic Review*, 75(1), 101–116.

Frey, B. S. (1997). *Not Just for the Money: An Economic Theory of Personal Motivation*. Edward Elgar Publishing.

Frey, B. S., & Jegen, R. (2001). Motivational interactions: Effects on behaviour. *Annales d'Économie et de Statistique*, 63–64, 131–153.

Gächter, S., & Thöni, S. (2010). Social comparison and performance: Experimental evidence on the fair-wage effort hypothesis. *Journal of Economic Behavior and Organization*, 76(3), 531–543.

Garretsen, H., Stoker, J. I., & Weber, R. A. (2020). Economic perspectives on leadership: Concepts, causality, and context in leadership research. *The Leadership Quarterly*, 31(3), 101410.

Gerhards, L., & Siemer, N. (2016). The impact of private and public feedback on worker performance – evidence from the lab. *Economic Enquiry*, 54(2), 1188–1201.

Gibbs, M. (1991). An economic approach to process in pay and performance appraisals. *Mimeo.*

Gill, D., Zdenka, K., Lee, J., & Prowse, V. (2019). First-place loving and last-place loathing: How rank in the distribution of performance affects effort provision. *Management Science*, 65(2), 494–507.

Gneezy, U., & Rustichini, A. (2000). Pay enough or don't pay at all. *Quarterly Journal of Economics*, 115(3), 791–810.

Greiner, B. (2015). Subject pool recruitment procedures: Organizing experiments with ORSEE. *Journal of the Economic Science Association*, 1(1), 114–125.

Hannan, R. L., Krishnan, R., & Newman, D. (2008). The effects of disseminating relative performance feedback in tournament versus individual performance compensation plans. *The Accounting Review*, 83(4), 893–913.

Harbring, C., & Irlenbusch, B. (2003). An experimental study on tournament design. *Labour Economics*, 10(4), 443–464.

Hölmstrom, B. (1979). Moral hazard and observability. *The Bell Journal of Economics*, 10(1), 74–91.

Hölmstrom, B., & Milgrom, P. (1991). Multitask principal–agent analyses: incentive contracts, asset ownership, and incentives. *Journal of Law Economics and Organization*, 7, 24–52.

Jackson, E. (2012). Ten biggest mistakes bosses make in performance reviews. *Forbes*, January 9, 2012. Available at: https://www.forbes.com/sites/ericjackson/2012/01/09/ten-reasons-performance-reviews-are-done-terribly/#642798e5ee07.

Judge, T. A., & Piccolo, R. F. (2004). Transformational and transactional leadership: A meta-analytic test of their relative validity. *Journal of Applied Psychology*, 89(5), 755–768.

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

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.

Kube, S., Marechal, A., & Puppe, C. (2013). Do wage cuts damage work morale – evidence from a natural field experiment. *Journal of European Economic Association*, 11(4), 853–870.

Kuhnen, C. M., & Tymula, A. (2012). Feedback, self-esteem, and performance in organizations. *Management Science*, 58(1), 94–113.

Lazear, E. P. (2000). Performance pay and productivity. *American Economic Review*, 90(5), 1346–1361.

Lizzeri, A., Meyer, M., & Persico, N. (2002). The incentive effects of interim performance evaluations. *CARESS*, Working Paper #02-09.

Locke, E. A., & Latham, G. P. (1990). *A Theory of Goal Setting and Task Performance*. Englewood Cliffs, NJ: Prentice-Hall.

Lonati, S., Quiroga, B. F., Zehnder, C., & Antonakis, J. (2018). On doing relevant and rigorous experiments: Review and recommendations. *Journal of Operations Management*, 64(1), 19–40.

Longenecker, C. O., Sims Jr, H. P., & Gioia, D. A. (1987). Behind the mask: The politics of employee appraisal. *The Academy of Management Executive*, 1(3), 183–193.

Lowe, K. B., Kroeck, K. G., & Sivasubramaniam, N. (1996). Effectiveness correlates of transformational and transactional leadership: A meta-analytic review of the MLQ literature. *The Leadership Quarterly*, 7(3), 385–425.

Mas, A., & Moretti, E. (2009). Peers at work. *American Economic Review*, 99(1), 112–145.

Mazar, N., Amir, O., & Ariely, D. (2008). The dishonesty of honest people: A theory of self-concept maintenance. *Journal of Marketing Research*, 45(6), 633–644.

Meslec, N., Curseu, P., Fodor, O. C., & Kenda, R. (2020). Effects of charismatic leadership and rewards on individual performance. *The Leadership Quarterly*, 31(6), 101423.

Murphy, K. R., & Cleveland, J. N. (1991). *Performance Appraisal: An Organizational Perspective*. Boston: Allyn and Bacoon.

Nae, E. Y., Moon, H. K., & Choi, B. K. (2015). Seeking feedback but unable to improve work performance? Qualified feedback from trusted supervisors matters. *Career Development International*, 20(1), 81–100.

Oswald, A. J., Proto, E. & Sgroi, D. (2015). Happiness and productivity. *Journal of Labor Economics*, 33(4), 789–822.

Peterson, S. J., & Luthans, F. (2006). The impact of financial and nonfinancial incentives on business-unit outcomes over time. *Journal of Applied Psychology*, 91(1), 156–165.

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.

Podsakoff, P. M., MacKenzie, S. B., & Bommer, W. H. (1996). Transformational leader behaviors and substitutes for leadership as determinants of employee satisfaction, commitment, trust, and organizational citizenship behaviors. *Journal of Management*, 22(2), 259–298.

Podsakoff, P. M., MacKenzie, S. B., Moorman, R. H., & Fetter, R. (1990). Transformational leader behaviors and their effects on follower's trust in leader, satisfaction, and organizational citizenship behaviors. *The Leadership Quarterly*, 1(2), 107–142.

Prendergast, C. (1999). The provision of incentives in firms. *Journal of Economic Literature*, 37(1), 7–63.

Prewitt, E. (2007). *Managing Performance to Maximize Results*. Harvard Business School Press.

Robson, A. J. (1992). Status, the distribution of wealth, private and social attitudes to risk. *Econometrica,* 60(4), 837–857.

Rosenthal, R., & Rosnow, R. L. (2009). *Artifacts in Behavioral Research: Robert Rosenthal and Ralph L. Rosnow's Classic Books: A Re-issue of Artifact in Behavioral Research, Experimenter Effects in Behavioral Research, and The Volunteer Subject*. New York: Oxford University Press.

Rustichini, A. (2008). Dominance and competition. *Journal of the European Economic Association*, 6(2–3), 647–656.

Sloof, R., & von Siemens, F. (2019). Effective leadership and the allocation and exercise of power in organizations. *The Leadership Quarterly,* forthcoming

Stajkovic, A. D., & Luthans, F. (2003). Behavioral management and task performance in organizations: Conceptual background, meta-analysis, and test of alternative models. *Personnel Psychology*, 56(1), 155–194.

Steffens, N. K., Fonseca, M. A., Ryan, M. K., Rink, F. A., Stoker, J. I., & Nederveen Pieterse, A. (2018). How feedback about leadership potential impacts ambition, organizational commitment, and performance. *The Leadership Quarterly*, 29(6), 637–647.

Szymanski, K., & Harkins, S. G. (1987). Social loafing and self-evaluation with a social standard. *Journal of Personality and Social Psychology*, 53(3), 891–897.

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

Vallacher, R., & Solodky, M. (1979). Objective self-awareness, standards of evaluation, and moral behavior. *Journal of Experimental Social Psychology*, 15(3), 254–262.

Winickoff, R. N., Coltin, K. L., Morgan, M. M., Buxbaum, R. C., & Barnett, G. O. (1984). Improving physician performance through peer comparison feedback. *Medical Care*, 22(6), 527–534.

Zehnder, C., Herz, H., & Bonardi, J. (2017). A productive clash of cultures: Injecting economics into leadership research. *The Leadership Quarterly, 28*(1), 65–85.

Zizzo, D. J. (2010). Experimenter demand effects in economic experiments. *Experimental Economics*, 13(1), 75–98.

Appendix for "Feedback quality and performance in organisations"

For online publication

Appendix A – Experimental instructions

[Note: These are the written instructions for the subjects facing the "No Feedback" treatment. The instructions for the "High Quality Feedback" and the "Low Quality Feedback" treatments appear in square brackets.]

INSTRUCTIONS

Welcome! You are about to take part in a decision-making experiment. This experiment is run by the "Birmingham Experimental Economics Laboratory" and has been funded by various research foundations. Just for showing up you have already earned £2.50. You will earn additional money during the experiment.

It is important that you remain silent and do not look at other people's work. If you have any questions, or need assistance of any kind, please raise your hand and an experimenter will come to you. If you talk, laugh, exclaim out loud, etc., you will be asked to leave and you will not be paid. We expect and appreciate your following of these rules.

We will first jointly go over the instructions. After we have read the instructions, you will have time to ask clarifying questions. Please do not touch the computer or its mouse until you are instructed to do so. Thank you.

In the instructions, unless otherwise stated, we will not speak in terms of British Pounds, but in terms of Experimental Currency Units (ECUs). Your entire earnings will, thus, be calculated in ECUs. At the end of the experiment the total amount of ECUs you have earned will then be converted into British Pounds at the following rate: **10 ECUs = £0.75**. The converted amount will be privately paid to you in cash at the end of the experiment.

Detailed Information about the Experiment

At the beginning of the experiment participants will be randomly matched into groups of four. The group composition will remain the same throughout the experiment. At no point during the experiment, nor afterwards will you be informed about the identity of the other participants in your group and the other participants will never be informed about your identity.

The experiment consists of 10 rounds during which you have to perform a task, as described below. At the beginning of each round you will receive 10 ECUs to perform the task. This payoff is independent of your performance during the task.

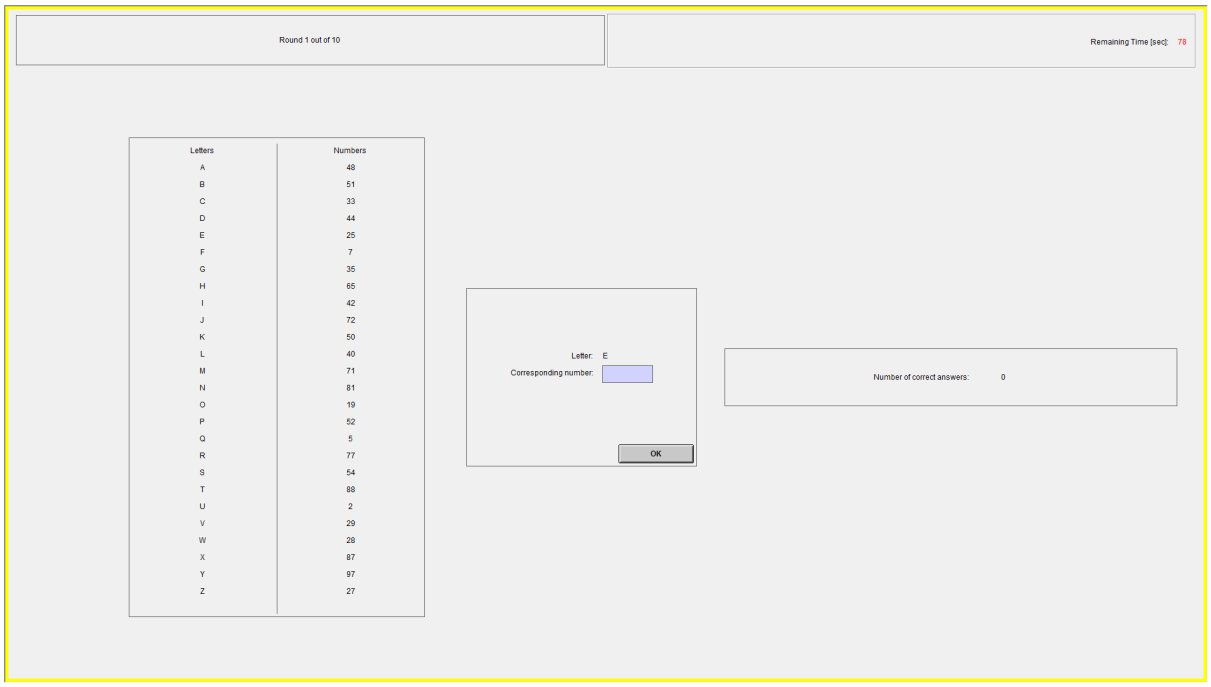
The Task

The task consists of converting letters into numbers during 90 seconds. Your screen displays a table with two columns. The first column indicates letters and the second column indicates their correspondence in numbers. You are given a letter and you must enter the corresponding number in the box on your screen. You must validate your answer by pressing the 'OK' button. Once you have validated your answer, you are immediately informed whether your answer is

correct or not. If your answer is incorrect, you must enter a new number until the answer is correct. A new letter appears only after you have submitted a correct answer for the current letter.

As soon you have validated a correct answer, the conversion table of letters and numbers is modified and a new letter to convert is displayed on your screen. You can convert as many letters as you like during the 90 seconds.

Below you can find a copy of the screenshot that will be displayed.



During each round, you are continuously informed of the remaining time until the end of the round (at the top right of the screenshot: ‘remaining time’) and of your score (the number of correct answers). In contrast, you will never be informed on the scores of the other three co-participants in your group.

At the end of the 90 seconds, you will be asked to guess what your performance expectation for that particular round is, relative to the other three co-participants in your group. In particular, you will have to choose a number from 1 (meaning that you are the first in your group) to 4 (meaning that you are the last in your group). If your guess is correct, you will receive £0.10 on top of your earnings from the task. On the same screen, you will also be asked to indicate how much effort you put in performing the task for that particular round on a 6-point scale (1 meaning "no effort at all" and 6 meaning "a lot of effort").

[*HQF and LQF treatments:* Following this, you will be provided with feedback about your performance relative to the other three co-participants in your group.]

Do you have any questions? Please raise your hand and an experimenter will come to your desk. Please do not ask any question out loud.

Appendix B – Additional analysis

Figure B.1. Distribution for the number of ranks experienced by subjects in the NF treatment

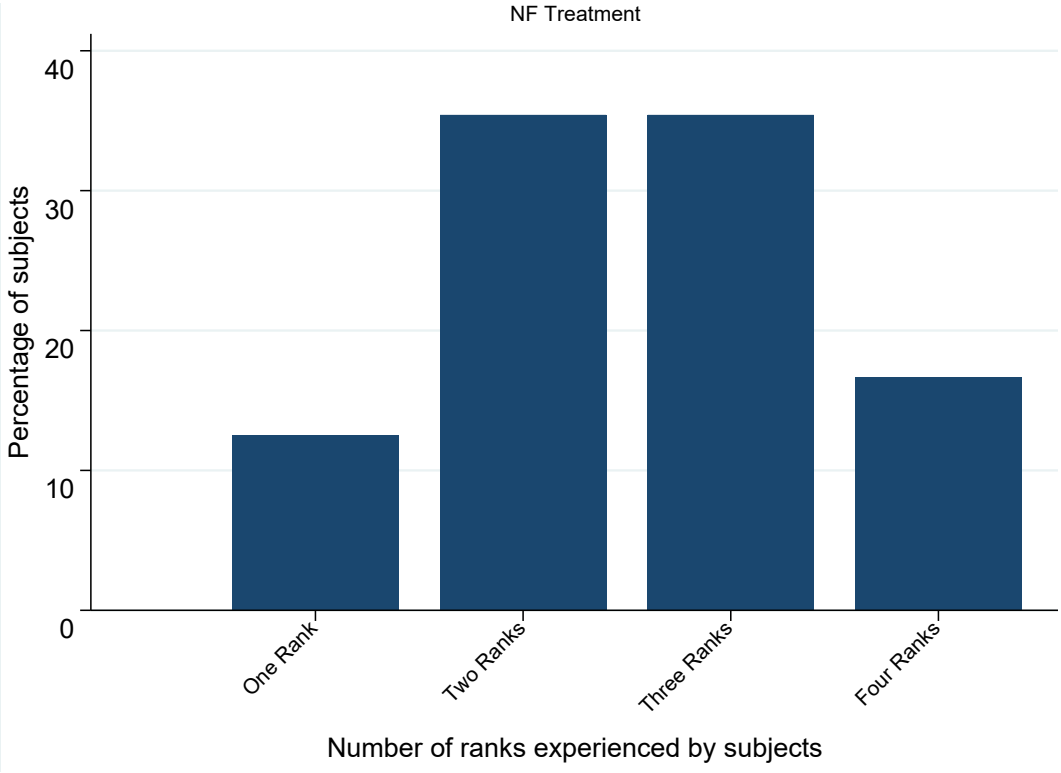


Table B.1. Pearson correlation coefficients of the reported variables across all treatments and in each treatment separately

Correlation Coefficients (All Treatments, N=248, 62 clusters)									
	M[a]	SD[a]	1	2	3	4	5	M[b]	SD[b]
1. Correct Answers	26.247	8.349		-0.674***	0.241†	-0.090	0.047	26.247	3.854
2. No Effort	0.049	0.163	-0.692***		-0.290*	0.020	0.191	0.049	0.074
3. Female	0.565	0.497	0.157*	-0.211***		0.226†	-0.150	0.565	0.239
4. Economics Degree	0.278	0.449	-0.001	0.063	0.019		-0.054	0.278	0.231
5. Nationality	0.665	0.473	-0.118†	0.160*	-0.054	-0.151*		0.665	0.217

Correlation Coefficients (NF Treatment, N=48, 12 clusters)									
	M[a]	SD[a]	1	2	3	4	5	M[b]	SD[b]
1. Correct Answers	23.731	10.02		-0.676*	0.217	-0.192	-0.265	23.731	4.76
2. No Effort	0.085	0.214	-0.689***		-0.134	0.027	0.432	0.085	0.09

3. Female	0.500	0.505	0.088	-0.128		0.243	-0.087	0.500	0.26
4. Economics Degree	0.208	0.410	-0.034	0.108	-0.000		0.000	0.208	.179
5. Nationality	0.625	0.489	-0.171	0.211	0.000	-0.133		0.625	0.250

Correlation Coefficients (LQF Treatment, N=96, 24 clusters)

	M[a]	SD[a]	1	2	3	4	5	M[b]	SD[b]
1. Correct Answers	25.73	8.38		-0.678***	0.331	-0.296	0.005	25.734	3.51
2. No Effort	0.057	0.17	-0.741***		-0.405*	0.099	0.330	0.057	0.07
3. Female	0.583	0.49	0.305**	-0.313**		0.261	-0.345†	0.583	0.25

4. Economics Degree	0.333	0.47	-0.093	0.087	0.015		-0.095	0.333	0.26
5. Nationality	0.656	0.47 7	-0.202*	0.245*	-0.211*	-0.186†		0.656	0.21 9

Correlation Coefficients (HQF Treatment, N=104, 26 clusters)

	M[a]	SD[a]	1	2	3	4	5	M[b]	SD[b]
1. Correct Answers	27.88	7.10 2		-0.509**	0.096	0.159	0.220	27.882	2.96
2. No Effort	0.024	0.12	-0.603***		-0.247	-0.059	-0.026	0.024	0.05
3. Female	0.577	0.49	0.028	-0.151		0.140	-0.009	0.577	0.22
4. Economics Degree	0.260	0.44	0.118	0.009	0.019		-0.043	0.260	0.21

5. Nationality	0.692	0.464	-0.026	0.046	0.062	-0.128		0.692	0.204

Notes. The variable "Correct Answers" corresponds to the number of correct answers provided by subjects in the encryption task. The variable "No Effort" corresponds to the percentage of subjects putting zero correct and zero wrong answers in the encryption task. "Female" is a dummy variable that takes the value 1 if a subject is female and 0 otherwise. "Economics degree" is a dummy variable which takes the value 1 if a subject is studying for an Economics/Business degree and 0 otherwise, and "Nationality" takes the value 1 if a subject comes from the UK or another European country and 0 otherwise.

[a] at individual level (with correlations below the diagonal)

[b] at matching group level (with correlations above the diagonal)

† $p < .10$

* $p < .05$

** $p < .01$

*** $p < .001$

Table B.2. Regressions of performance on linear round trend in each treatment

	Dependent variable: Number of correct answers		
Independent variables	NF	LQF	HQF
Rounds	-0.193	0.185†	0.409***
	(0.150)	(0.095)	(0.099)
Constant	24.794***	24.719***	25.635***
	(1.047)	(0.743)	(0.653)
Observations	480	960	1,040

Notes. OLS regressions. Robust standard errors, clustered at the matching group level (62 clusters), are reported in parentheses.

† $p < .10$

*** $p < .001$

Table B.3. Regressions of performance on period dummies in each treatment

	Dependent variable: Number of correct answers		
Independent variables	NF	LQF	HQF
Round 2	2.750**	3.896***	3.683***
	(0.646)	(0.596)	(0.437)
Round 3	3.146*	5.302***	4.644***
	(1.396)	(0.482)	(0.455)
Round 4	2.854**	4.771***	4.779***
	(0.713)	(0.612)	(0.669)
Round 5	0.604	4.448***	5.163***
	(1.508)	(0.717)	(0.765)
Round 6	1.333	4.625***	5.337***
	(1.657)	(0.632)	(0.747)
Round 7	1.292	4.813***	5.038***
	(1.591)	(0.627)	(0.925)
Round 8	1.000	3.938***	5.375***
	(1.727)	(0.725)	(0.840)
Round 9	−0.271	3.646***	5.423***
	(1.576)	(0.743)	(0.886)
Round 10	0.438	4.302***	5.625***
	(1.460)	(0.864)	(0.869)
Constant	22.417***	21.760***	23.375***
	(1.163)	(0.669)	(0.667)
Observations	480	960	1,040

Notes. OLS regressions. Robust standard errors, clustered at the matching group level (62 clusters), are reported in parentheses.

* $p < .05$

** $p < .01$

*** $p < .001$

Table B.4. Performance differences by treatment (controlling for demographic variables)

Dependent variable: Number of correct answers				
Independent variables	All Rounds	Rounds 1-5	Rounds 6-10	Rounds 2-10
LQF	1.931	0.982	2.881†	2.238
	(1.448)	(1.362)	(1.674)	(1.518)
HQF	4.139**	2.614†	5.665**	4.513**
	(1.406)	(1.330)	(1.618)	(1.472)
Rounds	0.205**	0.883***	-0.095	-0.025
	(0.068)	(0.130)	(0.100)	(0.068)
Constant	22.883***	20.989***	25.146***	24.484***
	(1.584)	(1.528)	(1.853)	(1.649)
Observations	2,480	1,240	1,240	2,232
Differences in average performance between treatments				
LQF minus HQF	-2.208*	-1.632†	-2.784*	-2.275*
	(0.917)	(0.857)	(1.097)	(0.955)

Notes. OLS regressions. Robust standard errors, clustered at the matching group level (62 clusters), are reported in parentheses.

† p < .10

* p < .05

** p < .01

*** p < .001

Table B.5. Performance differences by treatment (with period dummies)

Dependent variable: Number of correct answers				
Independent variables	All Rounds	Rounds 1-5	Rounds 6-10	Rounds 2-10
LQF	2.003	1.156	2.850	2.299
	(1.509)	(1.404)	(1.739)	(1.579)
HQF	4.150**	2.741*	5.560**	4.505**
	(1.451)	(1.338)	(1.687)	(1.523)
Round 2	3.585***	3.585***		
	(0.319)	(0.319)		
Round 3	4.609***	4.609***		1.024**
	(0.382)	(0.382)		(0.327)
Round 4	4.403***	4.403***		0.819*
	(0.397)	(0.397)		(0.398)
Round 5	4.004***	4.004***		0.419
	(0.548)	(0.548)		(0.490)
Round 6	4.286***			0.702
	(0.533)			(0.513)
Round 7	4.226***		−0.060	0.641
	(0.570)		(0.415)	(0.502)
Round 8	3.972***		−0.315	0.387
	(0.584)		(0.394)	(0.532)
Round 9	3.633***		−0.653	0.048
	(0.609)		(0.396)	(0.576)

Round 10	4.109***		−0.177	0.524
	(0.608)		(0.455)	(0.553)
Constant	20.049***	20.967***	23.416***	23.370***
	(1.274)	(1.199)	(1.509)	(1.311)
Observations	2,480	1,240	1,240	2,232
Differences in average performance between treatments				
LQF minus HQF	-2.147*	-1.585†	-2.710*	-2.207*
	(0.914)	(0.865)	(1.086)	(0.947)

Notes. OLS regressions. Robust standard errors, clustered at the matching group level (62 clusters), are reported in parentheses.

† $p < .10$

* $p < .05$

** $p < .01$

*** $p < .001$

Table B.6. Performance differences by treatment (controlling for demographic variables) with period dummies

Dependent variable: Number of correct answers				
Independent variables	All Rounds	Rounds 1-5	Rounds 6-10	Rounds 2-10
LQF	1.931	0.982	2.881†	2.238
	(1.451)	(1.364)	(1.676)	(1.520)
HQF	4.139**	2.614†	5.665**	4.513**
	(1.408)	(1.332)	(1.620)	(1.475)
Round 2	3.585***	3.585***		
	(0.319)	(0.319)		
Round 3	4.609***	4.609***		1.024**
	(0.382)	(0.383)		(0.327)
Round 4	4.403***	4.403***		0.819*
	(0.397)	(0.398)		(0.399)
Round 5	4.004***	4.004***		0.419
	(0.548)	(0.549)		(0.491)
Round 6	4.286***			0.702
	(0.533)			(0.514)
Round 7	4.226***		−0.060	0.641
	(0.570)		(0.415)	(0.502)
Round 8	3.972***		−0.315	0.387
	(0.585)		(0.394)	(0.533)

Round 9	3.633***		−0.653	0.048
	(0.609)		(0.396)	(0.576)
Round 10	4.109***		−0.177	0.524
	(0.609)		(0.455)	(0.553)
Constant	20.330***	20.317***	24.629***	23.829***
	(1.581)	(1.532)	(1.768)	(1.582)
Observations	2,480	1,240	1,240	2,232
Differences in average performance between treatments				
LQF minus HQF	-2.208*	-1.632†	-2.784*	-2.275*
	(0.919)	(0.858)	(1.099)	(0.956)

Notes. OLS regressions. Robust standard errors, clustered at the matching group level (62 clusters), are reported in parentheses.

† $p < .10$

* $p < .05$

** $p < .01$

*** $p < .001$

Table B.7. Performance differences by treatment (controlling for quadratic trend)

Dependent variable: Number of correct answers				
Independent variables	All Rounds	Rounds 1-5	Rounds 6-10	Rounds 2-10
LQF	2.003	1.156	2.850	2.299
	(1.507)	(1.403)	(1.738)	(1.577)
HQF	4.150**	2.741*	5.560**	4.505**
	(1.449)	(1.337)	(1.686)	(1.521)
Rounds	1.395***	4.824***	-1.224	0.229
	(0.185)	(0.351)	(1.125)	(0.243)
Rounds^2	-0.108***	-0.657***	0.071	-0.021
	(0.014)	(0.061)	(0.070)	(0.018)
Constant	20.223***	17.041***	28.308***	23.405***
	(1.296)	(1.247)	(4.490)	(1.434)
Observations	2,480	1,240	1,240	2,232
Differences in average performance between treatments				
LQF minus HQF	-2.147*	-1.585†	-2.710*	-2.207*
	(0.913)	(0.864)	(1.085)	(0.946)

Notes. OLS regressions. Robust standard errors, clustered at the matching group level (62 clusters), are reported in parentheses.

† $p < .10$

* $p < .05$

** $p < .01$

*** $p < .001$

Table B.8. Performance differences by treatment (controlling for quadratic trend and demographic variables)

Dependent variable: Number of correct answers				
Independent variables	All Rounds	Rounds 1-5	Rounds 6-10	Rounds 2-10
LQF	1.931	0.982	2.881†	2.238
	(1.449)	(1.363)	(1.675)	(1.518)
HQF	4.139**	2.614†	5.665**	4.513**
	(1.406)	(1.331)	(1.619)	(1.473)
Rounds	1.395***	4.824***	-1.224	0.229
	(0.185)	(0.351)	(1.126)	(0.243)
Rounds^2	-0.108***	-0.657***	0.071	-0.021
	(0.014)	(0.061)	(0.070)	(0.018)
Constant	20.504***	16.390***	29.521***	23.864***
	(1.618)	(1.564)	(4.858)	(1.731)
Observations	2,480	1,240	1,240	2,232
Differences in average performance between treatments				
LQF minus HQF	-2.208*	-1.632†	-2.784*	-2.275*
	(0.917)	(0.857)	(1.098)	(0.955)

Notes. OLS regressions. Robust standard errors, clustered at the matching group level (62 clusters), are reported in parentheses.

† $p < .10$

* $p < .05$

** $p < .01$

*** $p < .001$

Table B.9. Zero effort choices by treatment

Dependent variable: No Effort				
Independent variables	All Rounds	Rounds 1-5	Rounds 6-10	Rounds 2-10
LQF	-0.217	-0.265	-0.188	-0.228
	(0.218)	(0.260)	(0.243)	(0.216)
HQF	-0.615*	-0.538†	-0.661*	-0.652*
	(0.271)	(0.304)	(0.301)	(0.272)
Rounds	0.076***	0.095*	0.082***	0.076***
	(0.020)	(0.048)	(0.023)	(0.020)
Constant	-1.822***	-1.872***	-1.878***	-1.806***
	(0.205)	(0.248)	(0.287)	(0.204)
Observations	2,480	1,240	1,240	2,232
Differences in no effort between treatments				
LQF minus HQF	0.398†	0.272	0.473†	0.425†
	(0.240)	(0.302)	(0.253)	(0.240)

Notes. Probit regressions. Robust standard errors, clustered at the matching group level (62 clusters), are reported in parentheses.

† $p < .10$

* $p < .05$

*** $p < .001$

Table B.10. Balance checks of individuals' characteristics across treatments

	LQF treatment	LQF (reverse) treatment	Mann-Whitney test
	(1)	(2)	(3)
Female	0.58	0.48	0.151
	(0.49)	(0.50)	
Economics degree	0.33	0.25	0.216
	(0.47)	(0.43)	
Nationality	0.66	0.60	0.450
	(0.48)	(0.49)	
Obs.	96	88	

Notes: The table reports information on the students' demographic variables such as gender, field of study and nationality. "Female" is a dummy variable that takes the value 1 if a subject is female and 0 otherwise. "Economics degree" is a dummy variable which takes the value 1 if a subject is studying for an Economics/Business degree and 0 otherwise, and "Nationality" takes the value 1 if a subject comes from the UK or another European country and 0 otherwise. Columns 1-2 report average and standard deviations (in parentheses) of each demographic variable in the LQF and LQF (reverse) treatments. Column 3 reports two-sided p-values from a Mann-Whitney test.