What Is a Good Rank? The Effort and Performance Effects of Adding Performance Category Labels to Relative Performance Information*

THORSTEN KNAUER, *Ruhr University Bochum*

FRIEDRICH SOMMER, *University of Bayreuth*

ARNT WÖHRMANN, *Giessen University*†

ABSTRACT

Prior research demonstrates that relative performance information affects effort and performance. However, little is known about the qualitative design parameters of these information systems. This study examines, via an experiment, how adding performance category labels to ranks (e.g., "good" ranking position and "poor" ranking position) affects effort and performance. Furthermore, we investigate the effort and performance effects of two design choices observed in practice: the type of performance category labels and the proportion of positively labeled ranks. We argue that performance category labels motivate greater effort and performance through competition for status, which varies with both the type of performance category labels and the proportion of positively labeled ranks. We find partial support for our hypothesis that adding performance category labels increases effort and performance. Specifically, we find positive effects if top ranks are positively labeled and bottom ranks are negatively labeled (combined labels) but not if only top ranks are labeled (positive-only labels). We also find as predicted that the positive effects on effort resulting from using combined labels, instead of positive-only labels, are stronger when the proportion of positively labeled ranks is larger. The results for performance are weaker. Our results shed new light on the usefulness of performance category labels and emphasize how firms can render relative performance information more effective.

Keywords: relative performance, category labels, incentives, rankings, status, regulatory focus

Qu'est-ce qu'un bon classement? Les répercussions sur les efforts et le rendement de l'attribution d'étiquettes de catégorie de rendement aux données sur le rendement relatif

RÉSUMÉ

De précédentes recherches ont démontré que les données sur le rendement relatif influent sur les efforts et le rendement. Nous savons toutefois peu de choses au sujet des paramètres de conception

* Accepted by Khim Kelly. We thank Alan Webb (editor-in-chief), Khim Kelly (editor), two anonymous reviewers, Victor Maas, Michael Majerczyk, Adam Presslee, Anja Schwering, Marcel van Rinsum, Ivo Tafkov, and the participants and discussants at the University of Münster workshop; the Accounting Research Seminar at the University of Amsterdam; the 2015 Accounting, Behavior and Organizations Research Conference; the 2016 Management Accounting Section Midyear Meeting; and the European Accounting Association Annual Congress 2016 for their helpful comments and suggestions. We also appreciate the financial support provided by Dr. Werner Jackstädt-Stiftung and the Chartered Institute of Management Accountants (CIMA).

† Corresponding author.

Vol. 38 No. 2 (Summer 2021) pp. 839–866
doi:10.1111/1911-3846.12657

qualitative de ces systèmes d'information. Les auteurs procèdent à une expérience visant à déterminer comment l'attribution aux classements d'étiquettes de catégorie de rendement (pour un classement « bon » ou « faible », par exemple) influe sur les efforts et le rendement. Ils analysent en outre les répercussions sur les efforts et le rendement de deux choix de conception observés dans la pratique : la nature des étiquettes de catégorie de rendement et la proportion de classements auxquels est attribuée une étiquette de catégorie de rendement positive. Selon les auteurs, l'attribution d'étiquettes de catégorie de rendement motive une augmentation des efforts et du rendement dans le cadre d'une rivalité quant au statut, qui varie à la fois en fonction de la nature des étiquettes de catégorie de rendement et de la proportion de classements auxquels est attribuée une étiquette positive. Les données que recueillent les auteurs confirment en partie leur hypothèse selon laquelle l'attribution d'étiquettes de catégorie de rendement accroît les efforts et le rendement. Ils constatent plus précisément des répercussions positives lorsque les classements supérieurs ont une étiquette positive et les classements inférieurs, une étiquette négative (étiquettes combinées), ce qui n'est pas le cas lorsque des étiquettes ne sont attribuées qu'aux classements supérieurs (et qu'elles sont donc toutes positives). Ils observent également, conformément à leur hypothèse, que les répercussions positives sur les efforts de l'utilisation d'étiquettes combinées, plutôt que d'étiquettes uniquement positives, sont davantage marquées lorsque la proportion de classements auxquels est attribuée une étiquette positive est plus importante. Les résultats pour ce qui est du rendement sont plus mitigés. Les conclusions de l'étude jettent un nouvel éclairage sur l'utilité de l'attribution d'étiquettes de catégorie de rendement et sur la façon dont les entreprises peuvent accroître l'efficacité des données sur le rendement relatif.

Mots clés : rendement relatif, étiquettes de catégorie de rendement, stimulants, classements, statut, orientation régulatrice

1. Introduction

Firms frequently compare employees with their peers and rank them based on some type of performance measure (Hazels and Sasse 2008). Firms that follow this approach can provide fine relative performance information (RPI) or coarse ranking information.[1] Thus, employees can learn about either their individual rank or the broad performance category to which their rank belongs (coarse ranking). In coarse rankings, firms may use coarse performance category labels, such as "'top', 'good', 'fair', 'poor' . . . and so forth" (Lipman 2012). There are different reasons why firms choose one system or the other. For example, some banks rank tellers based on the exact number of accounts opened and provide RPI for this measure (Tafkov 2013). In other situations—for example, when firms want to measure other attributes, such as employees' initiative or creativity (Caruth and Handlogten 2001)—such precise measurement appears to be infeasible or too costly. The same is true when the number of employees is high, and precise rank ordering would be—at least to some extent—arbitrary (Murphy and Cleveland 1995).

RPI and coarse performance category labels can also be combined. This approach is particularly appealing if RPI is available anyway, and firms additionally divide the rank order into coarser categories at no or low cost. For example, Ryanair publicly ranks its pilots on fuel consumption (using fuel-league tables). The top 20 pilots are assigned to an (unrewarded) positive performance category and receive a letter of appreciation from the firm, while the bottom 20 pilots are assigned to a negative performance category and receive negative feedback (Carbery 2012; Seher 2012). For example, a pilot ranked #15 receives precise information about the ranking position and receives a letter of appreciation noting that the ranking position falls into the top performance category.

1. While coarse ranking information is also a type of relative performance information, this paper uses the term RPI to refer specifically to fine individual performance or fine individual rankings.

Firms that provide coarse performance category labels, in addition to RPI, must inevitably determine two design choices: the type of performance category labels and the proportion of positively labeled ranks. Regarding the type of performance category labels, a wide variety of labels exists in practice. Some firms appraise only the top performers, such as "employees of the month." In this case, only the top performance category is positively labeled, while all other ranks are unlabeled (*positive-only labels*).[2] Other firms, such as Ryanair, employ systems that combine both positive and negative categories (*combined labels*). For both positive-only labels and combined labels, firms must determine the proportion of ranks that fall into the positively labeled category, for example, whether the top 5% or the top 30% of ranked employees fall into this category.

Whether or not adding (uncompensated) performance category labels to RPI and the particular design features of these labels affect effort and performance is an open, yet important, question. We leverage social comparison and regulatory focus theories to predict that, in the presence of RPI, performance category labels, especially combined labels, lead to a higher effort than no labels. We argue that adding performance category labels to RPI intensifies social comparisons and that employees—particularly middle performers—start competing for status, complementing the competition for the ranks induced by RPI (Festinger 1954; Tesser 1988).[3] We expect effort to be greater for combined labels compared to positive-only labels because adding negative labels to the bottom ranks establishes a prevention focus of avoiding the shame associated with a negative label.[4] We make the same prediction for performance provided that the effort-performance link is sufficiently strong. Finally, we predict the positive effort and performance effects of using combined labels instead of positive-only labels to increase in the proportion of positively labeled ranks. The intuition behind this prediction is that belonging to the low-status group is more shameful when the low-status group is smaller.

We test our predictions via an experiment that uses a $2 \times 2 + 1$ between-subjects design. We hold constant that all participants learn about the ranking positions of all group members (public RPI). In the baseline (*no labels*) condition, no performance category labels are added. We manipulate the type of additional performance category labels available (*positive-only labels* vs. *combined labels*) and the proportion of positively labeled ranks (*low* vs. *high*). Notably, neither our label manipulation nor the proportion manipulation involves any monetary incentive.

The design of our experiment allows social comparisons to arise, which is important for our theory. Specifically, we employ a real-effort task that asks subjects to solve multiple-choice multiplication problems. We capture effort through time spent on solving problems and performance through the number of problems solved with different levels of difficulty (easy, medium, and hard problems) to capture the different strengths of links between effort and performance. Furthermore, we provide public RPI based on the absolute number of problems solved, which provides a greater opportunity for social comparisons than private RPI. Finally, we use a piece-rate scheme because it motivates effort, and differences in performance thus more likely depend on ability; RPI is, therefore, more informative regarding relative ability.

We find partial support for our hypothesis that adding performance category labels to RPI increases effort and performance. In detail, combined labels increase effort but not positive-only labels. Furthermore, the positive effect on effort of using combined instead of positive-only labels

2. Labeling only the bottom ranks would be the other option for partial labeling in a two-tier system. However, in addition to raising ethical concerns, negative-only labels lack practical relevance (Dohmen 2012).
3. Labels can also be used to communicate additional information about absolute performance levels. Therefore, an employee could be the best in the ranking group, but the entire group might perform poorly compared to other groups. In the homogeneous groups that we investigate, this information, however, is not relevant.
4. Using a two-tier structure enables us to clearly identify the effect of (potential) belief revision since only good and poor ranks are assigned, instead of additional moderate or acceptable ranks.

is stronger when the proportion of positively labeled ranks is greater. Our results are similar for performance, provided that the effort-performance link is sufficiently strong, which it is for the easy problems but not for the medium/hard problems.

From a theory perspective, we add to the stream of research that investigates the effect of RPI on performance. Previous research in this field has investigated, for example, whether and how the following affect performance: the precision of RPI (Hannan et al. 2008), the publicity of RPI (Tafkov 2013), the pay link of RPI (Newman and Tafkov 2014), rank ordering (Kramer et al. 2016), and the use of causal language (Loftus and Tanlu 2017). We contribute by analyzing whether adding performance category labels to RPI affects effort and performance and whether the type of performance category labels and the proportion of positively labeled ranks matter.

Prior research allows no clear inferences regarding the effectiveness of adding performance category labels to RPI. Specifically, in two studies, Bhattacharya and Dugar (2012, 2013) find that in a math task, having performance-based status labels (e.g., the Top Performer, the Top Three Performers) increases performance vis-à-vis having no performance label, which suggests that individuals care about receiving the status of a top performer. However, given that the performance-based status labels include both RPI information and a status label, these studies are not able to disentangle the effects of status labels from RPI information. Our study adds to this research by investigating the incremental effect of *adding* labels to RPI. As RPI is always present in our study, it allows us to test if performance category labels have an incremental effect.

Charness et al. (2014) report in a flat-wage setting that adding a positive performance symbol (gold medal) to the top rank and a negative performance symbol (donkey hat) to the bottom rank *decreases* performance. They argue that the decrease in performance is attributable to nonmonetary performance symbols crowding out intrinsic motivation and people perceiving that their social image would be weakened by a signal that they are willing to work for nonmonetary positive symbolic rewards. The most important difference between Charness et al. (2014) and our study is that, while the former finds a negative effect of combined performance category labels on performance, we find a positive effect. Our setting differs from Charness et al. (2014) because our piece-rate incentive scheme motivates individuals to exert effort and to care about their performance. Thus, performance differences are likely attributed to differences in ability (Tafkov 2013), and good performance fosters one's social image. Therefore, our study complements Charness et al. (2014) by showing that adding combined performance category labels to RPI increases effort and performance when employees can draw conclusions regarding their relative ability, which is important for the development of social comparisons.

We also add to the literature by investigating the different effects of positive-only and combined labels, given that the related research has focused on either positive-only labels (Bhattacharya and Dugar 2012, 2013) or combined labels (Charness et al. 2014). We show that effort and performance increase more when combined labels are used compared to positive-only labels.

Finally, we contribute to the literature by investigating the interaction effect of the type of performance category labels and the proportion of positively labeled ranks. Previous research in this field has focused on how providing a greater proportion of employees with positive, unrewarded symbols affects performance (Bhattacharya and Dugar 2012, 2013; Moldovanu et al. 2007). We show that the effectiveness of using combined labels, instead of using positive-only labels, depends on the proportion of positively labeled ranks.

Our study has practical implications for firms that already provide RPI. We show that these firms can increase effort and performance if they add combined labels but not if they use positive-only labels. However, while the case of Ryanair illustrates that firms use combined labels, other firms prefer positive-only labels, for example, "employee(s) of the month" (Peterson and Luthans 2006). Our additional analysis provides a possible explanation: Employees perceive combined labels as less fair than positive-only labels, which may explain why firms commonly

use positive-only labels rather than combined labels.[5] Therefore, our study suggests that firms need to trade off the opportunity costs of foregone effort when using either no labels or positive-only labels instead of combined labels and the potential benefits of incentive schemes perceived as fairer by employees. Finally, we inform firms that are already providing rank RPI and performance category labels of the importance of both the type of label category used and the proportion of positively labeled ranks.

2. Development of hypotheses

Effects of performance category labels

According to economic theory, RPI should not affect effort or performance when compensation is independent of peer performance (Frederickson 1992). However, building on psychology research, previous studies document positive performance effects of RPI (Hannan et al. 2008, 2013; Kramer et al. 2016; Tafkov 2013). These studies attribute this effect to greater social comparison when RPI is available. Social comparison theory argues that employees have an innate drive to compare their performance to the performance of others to evaluate their own abilities (Festinger 1954). These comparisons directly affect one's self-image (Tesser 1988). Since individuals strive to maintain a positive self-image, performing better than others is important and induces positive feelings, such as pride, while negative feelings result from worse performance compared to others and leads to shame (Lazarus 1991; Smith 2000). Therefore, RPI allows employees to engage in social comparisons and motivates greater effort and performance through competition for *ranks* because employees attempt to achieve good ranking positions.

When performance category labels are added to RPI, a second benchmark for social comparisons in addition to RPI is introduced that works through the same mechanism as outlined by social comparison theory. As with RPI, performance category labels do not have to be monetarily rewarded to be effective. More precisely, performance category labels establish a good or bad status for a certain proportion of ranking positions. Status in this sense "is characterized by a rank-ordered relationship among people associated with prestige and deference behavior" (Huberman et al. 2004, 103; Ridgeway and Walker 1995). If employees' utility function incorporates a preference for status, they will accept the costs of effort to be rewarded with higher status. In fact, prior research finds that individuals care about status regardless of whether it comes with monetary benefits (Ball and Eckel 1998; Charness et al. 2014), and that firms can use this concern to increase performance (Besley and Ghatak 2008). When performance category labels are added, and thus the status of ranks is defined, social comparisons are intensified. Employees now also compete for (higher) *status* and attempt to achieve (avoid) a good (bad) status label.

Although competition for status affects performers at all ranks, performance category labels are more consequential for those individuals at the middle ranks. Middle performers who have barely achieved or just missed achieving a good status label face a greater risk of losing or a greater chance of winning a good status label. Furthermore, middle performers who have barely avoided or just received a bad status label have greater chances of avoiding a bad status label. Therefore, these labels can be particularly helpful in increasing the motivation of middle performers through competition for status.

In real-world situations, firms use various types of performance categories. Thus, we are also interested in whether the type of category label matters, and we compare positive-only labels and combined labels. We focus on a simple, two-tier system to examine whether labeling the bottom ranks also has a general effect. A two-tier, combined labels system differentiates between top

5. A similar argument is made in the literature that explores bonus versus penalty contracts (Hannan et al. 2005). Although penalty contracts result in greater effort, firms prefer bonus contracts that employees perceive as fairer. This is because employees would accept penalty contracts only if compensation was increased resulting in higher costs for the firm.

performers and all other performers by positively labeling the top ranks and negatively labeling all other ranks. Accordingly, combined and positive-only labels are similar in that top performers are assigned to a positively labeled, prestigious status group in both cases. However, they differ with respect to the treatment of the bottom performers. The bottom ranks are negatively labeled when combined labels are used, while these ranks are unlabeled in the case of positive-only labels.

We use regulatory focus theory (Higgins 1997, 1998) to predict that combined labels lead to more social comparisons than positive-only labels and thus lead to greater effort and performance. Regulatory focus theory differentiates between two separate self-regulation systems for a desired end: a promotion focus and a prevention focus (Higgins 1997, 1998). A promotion or prevention focus can be induced by both stable individual differences and situational factors (Förster et al. 1998; Shah and Higgins 2001). We focus on situational factors and argue that the presence of performance category labels is a situational factor that induces a promotion or prevention focus. Specifically, a goal can be framed in a way that motivates an employee to either attain a specific achievement (promotion goal) or prevent a specific failure (prevention goal). For example, an achievement goal would be recognition as being a member of a high-status group, while a prevention goal would be *avoiding* recognition as being part of a low-status group. According to regulatory focus theory, framing a goal as an achievement goal establishes a promotion focus, while framing a goal as a prevention goal establishes a prevention focus. Pennington and Roese (2003) use the example of an Olympic athlete to illustrate the difference between promotion and prevention goals. When the athlete has a promotion focus, he or she strives "toward getting onto the medal stand," while in the case of a prevention focus, he or she strives "to preserve the honor of his nation by not scoring in the bottom half" (Pennington and Roese 2003, 564). Importantly, promotion and prevention foci are not mutually exclusive but influence behavior separately from one another (Higgins 2002). This outcome is supported by the finding in neuroscience research that the activation of promotion versus prevention goals is associated with different brain regions (Eddington et al. 2007). Therefore, "one focus, both foci, or neither focus" can be present in an individual (Johnson et al. 2010, 232). According to regulatory focus theory, the presence of promotion and prevention foci leads to emotional reactions of different magnitudes and natures, resulting in different levels of motivation (Brockner and Higgins 2001). Since both foci influence behavior separately, regulatory focus theory predicts that the presence of both foci leads to greater motivation than the presence of either a promotion or a prevention focus.

The two types of labels differ regarding whether they establish *only* a promotion focus (positive-only labels) or *also* a prevention focus (combined labels). Since the difference is only whether a prevention focus is established, we examine the effects caused by adding a prevention focus. Based on regulatory focus theory, a prevention focus likely materializes if the bottom ranks are negatively labeled but not if the bottom ranks are unlabeled. If combined labels are used and induce *both* a promotion *and* a prevention focus, employees are likely motivated to both attain a positive label and prevent a negative label. If positive-only labels are used and establish only a promotion focus, employees will only be motivated to attain a positive label. Thus, more brain regions are activated when combined labels are used (Eddington et al. 2007), stronger reactions arise, and greater motivation develops if combined labels are used than if the bottom ranks are unlabeled. Therefore, combined labels weigh more heavily toward being explicitly known as a member of the high- versus the low-status group, while positive-only labels create the potential of being known only as a member of the high-status group.

Accordingly, the presence of performance category labels intensifies social comparison, and employees—particularly middle performers—become more competitive and start competing for status. Regarding the type of label, competition for status is greater when negative labels are added to the bottom ranks because a prevention focus is established, *in addition* to a promotion focus. Greater competition leads to greater effort. Therefore, we predict that, in the presence of RPI, combined labels lead to greater effort than positive-only labels, while positive-only labels

lead to greater effort than no labels. Hypothesis 1 is formally stated below and predicts the same pattern for performance as for effort, provided that the effort-performance link is sufficiently strong. If, however, the effort-performance link is weakened such as when performance is more dependent on other factors, particularly ability, the effect of performance category labels would weaken.

HYPOTHESIS 1. *If employees receive RPI, their (i) effort and (ii) performance are lowest when the ranks are unlabeled, higher if positive-only labels are used, and highest if combined labels are employed.*

Effects of the proportion of performance categories

Hypothesis 2 examines whether the difference between the types of labels depends on the proportion of ranks associated with a certain performance category label. This interaction is important because firms must determine the proportion of ranks with positive labels, that is, the size of the high-status group, regardless of whether only the top ranks are labeled (positive-only labels) or all ranks are labeled (combined labels). Firms define the proportion of bottom ranks simultaneously: either unlabeled *or* negatively labeled. Below, we argue that the positive effect on effort and performance of using negative labels for the bottom ranks in the combined labels condition, instead of no labels for the bottom ranks in the positive-only condition, is greater when the proportion of positively labeled ranks is larger.

Greater proportions of positively labeled ranks automatically lead to lower proportions of bottom ranks in the two-tier system. If a prevention goal is established, this greater proportion will emphasize the prevention goal. Thus, achieving this goal becomes more valuable because social comparisons make it increasingly difficult for an employee to rationalize not having attained a positively labeled rank or not belonging to the high-status group when a larger number of peers are successfully in the high-status group (Knauer et al. 2017). Regulatory focus theory predicts that, as "the value of a prevention goal increases, the goal becomes a necessity. . . . When a goal becomes a necessity, one must do whatever one can to attain it regardless of the ease or likelihood of goal attainment" (Higgins 1998, 35). That is, avoiding being explicitly known as a member of the low-status group becomes more valuable as the size of the high-status group grows (or the number of members in the low-status group decreases).[6]

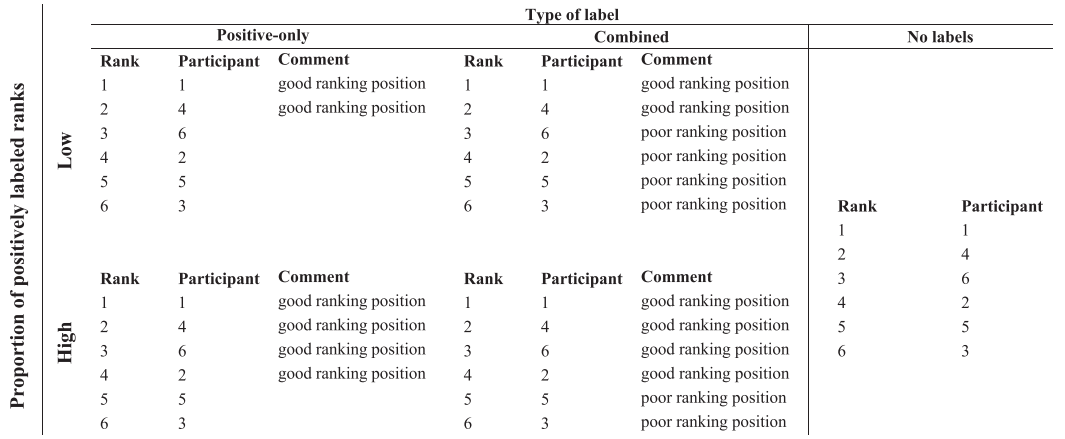
A prevention focus is established only when the bottom ranks are negatively labeled. Therefore, the pressure to avoid a bottom rank increases with a lower proportion of bottom ranks under combined labels; it arises to a smaller degree with positive-only labels. We therefore predict that the motivational effect of using combined labels, instead of positive-only labels, is stronger when the low-status group is smaller. This interaction effect of the type of label category and the proportion of positively labeled ranks is formally stated in Hypothesis 2. Again, we expect the effort effects to translate into performance effects if the effort-performance link is sufficiently strong.

HYPOTHESIS 2. *The positive (i) effort and (ii) performance effects of using combined labels, instead of positive-only labels, are stronger when the proportion of positively labeled ranks is larger.*

We do not include a prediction of a main effect of the proportion of positively labeled ranks because, as the proportion increases, two opposing effects could materialize at once. On the one

6. However, if the proportion of negatively labeled ranks becomes minuscule and employees feel certain that they will receive a positively labeled ranking position, complacency might set in and start reducing the positive effect of positively labeled ranks (Berger et al. 2013).

Figure 1 Illustration of experimental manipulations



hand, the top performance category becomes less scarce and therefore potentially loses its prestige, resulting in lower motivation. On the other hand, attaining the top performance category label appears to be more likely; thus, lower-performing individuals are willing to invest more effort to attain it (Knauer et al. 2017). The motivational effect of a reduced proportion of bottom ranks depends—as discussed above—on whether these are labeled. Taken together, we refrain from predicting a main effect.

3. Experimental method

Experimental design and manipulations

We use a $2 \times 2 + 1$ between-subjects experimental design.[7] The first factor is the type of performance category labels (*positive-only labels* vs. *combined labels*). The second factor is the proportion of positively labeled ranks (*low* vs. *high*). In the additional *no labels* condition, no performance category labels are provided; thus, the proportion of positively labeled ranks is not manipulated. In all treatment conditions, the participants are informed about the ranking positions of all group members (public RPI). The participants compete against one another in groups of six.

We manipulate performance category labels at two levels. In the positive-only labels conditions, the top ranks are labeled "good ranking position," and the bottom ranks are unlabeled. In the combined labels conditions, the top ranks are labeled "good ranking position," and the bottom ranks are labeled "poor ranking position." Therefore, in the combined labels condition, every rank is labeled. We vary the proportion of positively labeled ranks at two levels: low and high. In the low conditions, the two top ranks of six ranks have positive labels. In the high conditions, the top four ranks of six ranks have positive labels. The remaining ranks are either unlabeled (positive-only labels) or negatively labeled (combined labels).

Figure 1 illustrates the experimental manipulations for the five experimental conditions.

Our design follows the requirements suggested by Bhattacharya and Dugar (2012) for successful status induction via performance category labels. These requirements are: (i) labels are linked to performance; (ii) effort is costly; (iii) labels are publicly awarded; (iv) labels are

7. The task was programmed using z-Tree software (Fischbacher 2007). The research was conducted in an ethical manner. Specifically, subjects were treated anonymously in accordance with data protection regulations and were not exposed to specific risks. Furthermore, they were not deceived by any means and at any time and were debriefed via e-mail after the last session of the experiment was conducted. The institution where the study was conducted does not have a review board to provide ethics clearance.

specific; and (v) monetary rewards for good performance are provided, although performance category labels must be unrewarded.

We address these requirements by the following. First, we assign the participants to performance categories based on their (relative) performance. This link increases the credibility of the performance category label that supports status induction. It also requires homogenous participants, which we ensure through our participant selection and assignment process. If the groups of participants were too heterogeneous and, for example, one group consisted only of low (high) performers, lending a positive (negative) label to some of them would harm label credibility. Subjects were recruited from a relatively homogeneous student population and were randomly assigned to treatments, which they knew. Hence, they had no reason to assume heterogeneous abilities. Second, we make effort costly, as further explained in the section on the incentive scheme below. Third, we provide both performance category labels and RPI in public.[8] Fourth, we use unambiguous wording for the performance categories (i.e., "good" vs. "poor" ranking position). Finally, we provide piece-rate compensation with performance categories being unrewarded.

Experimental task

The experiment consisted of three trials in which the performance from a prior trial did not carry over to the next trial. Thus, *trial* was a within-subjects factor with three levels. Each trial consisted of three rounds. Using three rounds in each of the three trials increased the variance in the participants' performance, facilitating the assignment of ranks. Each round lasted up to 300 seconds. A visible clock on the computer screen, which counted down from 300 seconds, helped the participants to keep track of time. The participants could choose to stop working on the problems at any time by clicking the "Proceed to the next round" button.

In each of the three rounds of a trial, the subjects solved up to six multiplication problems (two easy, two medium, and two hard problems in each round) without using a calculator, a pen, or paper (Tafkov 2013). For each problem, five possible solutions were displayed. The subjects had to identify the correct solution. The problems were more difficult if more digits needed to be multiplied and if eliminating possible answers was more difficult. The participants knew that everyone received the identical problems in the same order and were free to choose the number of problems to work on and the order in which to work on them.

We used three difficulty levels and employed a multiple-choice task instead of requiring the participants to enter the correct solution, to allow the participants to draw more distinct conclusions regarding their relative ability. This is important for the development of social comparisons (Festinger 1954), which is central to our theory of how the status effects caused by performance category labels foster effort and performance. Specifically, using multiplication problems with different levels of difficulty decreases the marginal utility of effort when the subjects start working on harder problems since these require more time. More difficult problems also require more ability, while solving relatively easy problems primarily requires effort. In other words, the effort-performance link is stronger for easy problems and weakens for medium and hard problems. Since we limited the number of easy problems, the participants had reason to believe that performance differences likely depend on differences in ability and *not* only on differences in effort. Consequently, performance differences were more informative about relative ability because the participants perceived the multiplication task as informative regarding their general problem-solving skills. The multiple-choice task also allowed us to employ more challenging multiplication problems in which the participants had to find shortcuts to identify the correct solution,

8. In the real world, it appears to be sufficient if employees share common knowledge about which ranks fall into which category. For example, when Ryanair publicly ranks its pilots on fuel consumption, it is common knowledge that pilots at the top of the list receive a letter of appreciation, although this information is not explicitly included in the public performance ranking (Seher 2012).

instead of (only) conducting mechanical computations.[9] This technique further increased the informativeness of performance differences regarding the participants' relative ability.

At the end of each trial, public RPI was provided to the participants within each condition. We used public instead of private RPI because it provides greater opportunity for social comparisons and thus leads to greater feelings of pride and shame (Tafkov 2013). The participants were assigned ranks based on the number of correctly solved multiplication problems. The participants who solved more problems correctly received a better ranking position. The participants were informed that, if two participants had the same number of correctly solved problems, the two ranks in question would be assigned randomly.

Incentive scheme and dependent variables

All of the participants received 1,000 lira (the experimental currency) for showing up.[10] The participants could earn additional compensation if they clicked the "Proceed to the next round" button before the end of a round. Specifically, the participants earned a time bonus for completing a round before the 300 seconds were over—2 lira for every second saved. This mechanism is a common practice in accounting research (e.g., Hannan et al. 2008; Hecht et al. 2012; Knauer et al. 2017; Sprinkle 2000; Tafkov 2013). It ensured that the effort was costly, that is, that criterion (ii) from Bhattacharya and Dugar (2012) was fulfilled. Therefore, the *Time spent* working on the multiplication problems was our proxy for effort.[11]

In addition, the participants received 30 lira for each mathematical problem that they solved correctly (piece-rate payment) in all treatment conditions. Thus, compensation depended on the *absolute* number of correct answers, while the ranking position depended on *relative* performance compared to the group. We used a piece-rate scheme because prior research has shown that the informativeness of RPI and thus social comparison are greater if compensation is tied to performance, which is important for our theory (Tafkov 2013). This design feature is also consistent with criterion (v) in Bhattacharya and Dugar (2012) to render performance category labels more effective. The number of *Problems solved correctly* was our measure of performance.

Procedures

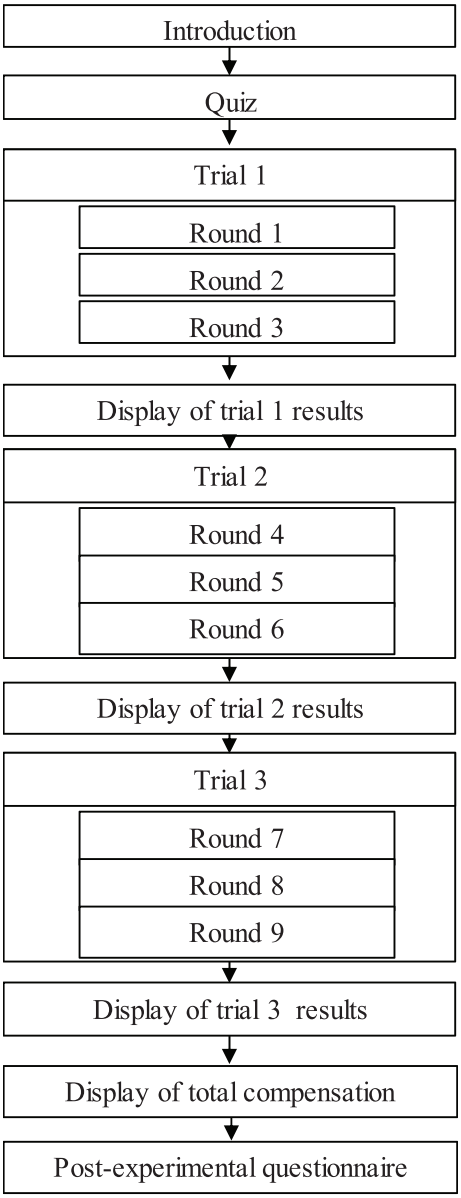
The experimental procedure is summarized in Figure 2. Upon arrival at the laboratory, the participants were given 10 minutes to read the written instructions. The instructions contained the experimental procedures and informed the subjects that they belonged to a group of six people competing against one another in three trials and that performance from earlier trials would not carry over to the subsequent trials. The subjects were informed that the groups were fixed throughout the entire experiment. Subsequently, the subjects were required to take a quiz to demonstrate an adequate understanding of the experimental procedures and the computation of their compensation. Specifically, they had to confirm their understanding that compensation depended *not* on ranks or labels but on the number of correct solutions and the bonus for the time saved. The participants were required to introduce themselves by standing up and stating their participant number (1–6) to ensure that they knew exactly against whom they would be competing (e.g., "I am participant #1"). Each computer had the corresponding participant number on top of the

9. One example of a shortcut for an easy problem is that, when two 2-digit numbers are multiplied, looking at the last digit allows for eliminating some incorrect answers. Participants were not told that they could find shortcuts to identify the correct answer.

10. At the end of the experiment, lira earned by participants was converted into euros at a rate of 380 lira per euro.

11. According to Baiman (1982), a proxy for effort must be controllable by the employee, be correlated with performance, and be costly. If the time spent working on the problems was not costly, all of the subjects would likely have taken the entire 300 seconds per round, but it would be unclear whether they were spending effort on solving the problems or were only waiting (perhaps for insight into solving the problem). Furthermore, making time costly reduced the possibility that the participants spent time on the problems because they enjoyed the task since that enjoyment would come at a cost.

Figure 2 Experimental procedures



monitor, and the participants were asked to put on T-shirts printed with their numbers when entering the laboratory.

After solving the multiplication problems in each of the three trials (which consisted of three rounds each), the participants were informed of their rank, the number of problems that they had solved correctly, and their compensation for the trial. The achieved rank was illustrated in a table. This table displayed which participant (indicated by the announced participant numbers) had achieved which rank (public RPI). The table also contained the experimental manipulation.

When all three trials were completed, the subjects received information about their total compensation. On average, the participants earned 11.82 euros for approximately 1.5 hours of participation. The participants were dismissed after completing a postexperimental questionnaire.

Participants

The participants were 150 undergraduate business students from a large Western European university; 30 were randomly assigned to each treatment condition. The average age of the participants was 20.5 years, and 83 (55.3%) were male. Given that our task required mathematical and general problem-solving skills, we asked the participants about the number of college-level math classes that they had attended. The subjects also judged their general problem-solving skills compared to the other members of their group on a scale from 1 (inferior) to 11 (superior). The subjects had completed 2.53 math classes on average and assumed their problem-solving skills to be slightly greater than the mean (7.40). Since there were no significant differences across conditions for age, gender, math background, or problem-solving skills (all p-values > 0.10, two-tailed), we conclude that the randomization was successful. Furthermore, the participants were sufficiently homogeneous across conditions regarding task-relevant characteristics, which was important for the label credibility, as explained above.

4. Results

Descriptive statistics

Table 1, panel A, reports the descriptive statistics for the two main dependent variables, *Time spent* and *Problems solved correctly*. On average, the participants dedicated 1,260.09 seconds to the multiplication problems out of a maximum of 2,700 seconds over all three trials and solved 20.33 problems correctly. Therefore, the participants spent, on average, 62 seconds to find a correct solution. As such, the participants deviated from a pure wealth-maximizing strategy because the compensation scheme applied paid 30 lira for each correct answer, while each second saved resulted in 2 lira. Since wealth-maximizing individuals would invest only up to 15 seconds to solve a problem, the participants were willing to forfeit compensation to engage in social comparisons.[12] Table 1, panel A, also shows that effort decreases over the three trials from 522.87 (rounds 1–3) to 397.47 (rounds 4–6) and finally to 339.75 (rounds 7–9) seconds. In contrast, performance increases from 5.99 (rounds 1–3) to 7.07 (rounds 4–6) and finally to 7.27 (rounds 7–9) correctly solved problems. These opposing trends indicate that problem-solving efficiency increases over the three trials.

Table 1, panel B, shows the descriptive data by performance category label when we pool the two manipulations of the proportions of positively labeled ranks. We use these data to test Hypothesis 1. Table 1, panel C, provides detailed descriptive data by performance category label for the number of problems solved by the level of difficulty. Additional analysis (untabulated) shows that, on average, over all treatments, identifying a correct solution for an easy problem takes 35 seconds, a medium problem takes 78 seconds, and a hard problem takes 108 seconds—with all pairwise comparisons significant at the 1% level in a two-tailed test.

Hypotheses tests

Hypothesis 1 predicts that, if employees receive RPI, their (i) effort and (ii) performance are lowest when the ranks are unlabeled, higher if positive-only labels are used and highest if combined labels are employed. We first discuss our results for effort. The descriptive statistics in Table 1, panel B, reveal that, in the presence of positive-only labels, effort increases from

12. To ensure that the participants were aware of this trade-off, two questions were asked on the quiz administered prior to the trials commencing, one directly after the other, to increase the salience of this information. The first question asked about the compensation for providing a correct solution, and the next question asked for the time bonus per second.

TABLE 1
Descriptive statistics

Panel A: Descriptive statistics (mean, [SD])

	Label							
		Low proportion			High proportion			
	No labels	Positive-only	Combined	Total	Positive-only	Combined	Total	Total
Number of subjects	30	30	30	60	30	30	60	150
Time spent								
Rounds 1–3 (Trial 1)	454.33	498.80	567.70	533.25	465.37	628.13	546.75	522.87
	[324.90]	[312.20]	[258.97]	[286.50]	[315.88]	[278.29]	[306.34]	[302.37]
Rounds 4–6 (Trial 2)	319.20	336.97	467.80	402.38	360.07	503.33	431.70	397.47
	[289.82]	[287.77]	[302.89]	[300.25]	[298.13]	[294.55]	[302.57]	[300.02]
Rounds 7–9 (Trial 3)	279.07	299.33	379.10	339.22	270.93	470.33	370.63	339.75
	[283.41]	[303.05]	[280.24]	[292.17]	[263.69]	[309.27]	[302.16]	[294.47]
Total	1,052.60	1,135.10	1,414.60	1,274.85	1,096.37	1,601.80	1,349.08	1,260.09
	[852.18]	[847.30]	[803.26]	[830.59]	[847.47]	[825.08]	[867.51]	[851.15]
Problems solved correctly								
Rounds 1–3 (Trial 1)	5.07	5.47	6.63	6.05	5.63	7.17	6.40	5.99
	[3.64]	[3.21]	[2.61]	[2.96]	[3.48]	[3.47]	[3.53]	[3.35]
Rounds 4–6 (Trial 2)	5.77	6.30	7.73	7.02	7.30	8.23	7.77	7.07
	[4.23]	[3.97]	[3.79]	[3.91]	[3.99]	[3.23]	[3.63]	[3.91]
Rounds 7–9 (Trial 3)	6.83	5.83	8.47	7.15	6.57	8.63	7.60	7.27
	[5.23]	[4.68]	[5.21]	[5.09]	[4.97]	[4.83]	[4.97]	[5.04]
Total	17.67	17.60	22.83	20.22	19.50	24.03	21.77	20.33
	[12.34]	[10.48]	[10.03]	[10.51]	[11.84]	[10.04]	[11.12]	[11.16]

(The table is continued on the next page.)

TABLE 1 (continued)

Panel B: Descriptive statistics by performance category label (mean, [SD])

	No labels	Positive-only	Combined	Total
Number of subjects	30	60	60	150
Time spent	1,052.60	1,115.74	1,508.20	1,260.09
	[852.18]	[840.40]	[812.81]	[851.15]
Problems solved correctly	17.67	18.55	23.43	20.33
	[12.34]	[11.13]	[9.97]	[11.16]
Easy problems solved correctly	9.00	10.38	12.88	11.11
	[6.49]	[5.28]	[4.65]	[5.49]
Medium problems solved correctly	5.67	5.20	7.23	6.11
	[5.14]	[4.73]	[4.60]	[4.82]
Hard problems solved correctly	3.00	2.97	3.32	3.11
	[2.10]	[2.67]	[2.68]	[2.56]

Panel C: Number of problems solved by level of difficulty (mean, [SD])

		Label						
		Low proportion			High proportion			
	No labels	Positive-only	Combined	Total	Positive-only	Combined	Total	Total
Number of subjects	30	30	30	60	30	30	60	150
Easy problems solved correctly								
Rounds 1–3 (Trial 1)	2.60	3.20	3.90	3.55	3.07	3.93	3.50	3.34
	[1.94]	[1.97]	[1.58]	[1.81]	[1.89]	[1.89]	[1.93]	[1.91]
Rounds 4–6 (Trial 2)	3.17	3.70	4.40	4.05	3.97	4.73	4.35	3.99
	[2.51]	[2.05]	[1.83]	[1.96]	[1.94]	[1.70]	[1.85]	[2.07]
Rounds 7–9 (Trial 3)	3.32	3.50	4.30	3.90	3.33	4.50	3.92	3.77
	[2.40]	[2.29]	[2.14]	[2.23]	[2.22]	[2.00]	[2.17]	[2.24]
Total	9.00	10.40	12.60	11.50	10.37	13.16	11.77	11.11
	[6.49]	[5.44]	[4.76]	[5.19]	[5.22]	[4.60]	[5.08]	[5.49]

(The table is continued on the next page.)

TABLE 1 (continued)

Panel C: Number of problems solved by level of difficulty (mean, [SD])

		Label						
		Low proportion			High proportion			
	No labels	Positive-only	Combined	Total	Positive-only	Combined	Total	Total
Medium problems solved correctly								
Rounds 1–3 (Trial 1)	1.60	1.47	1.87	1.67	1.53	2.20	1.87	1.73
	[1.65]	[1.50]	[1.50]	[1.50]	[1.72]	[1.58]	[1.67]	[1.60]
Rounds 4–6 (Trial 2)	1.90	1.83	2.33	2.08	2.00	2.70	2.35	2.15
	[1.81]	[1.68]	[2.04]	[1.87]	[1.82]	[1.70]	[1.78]	[1.82]
Rounds 7–9 (Trial 3)	2.17	1.77	2.70	2.23	1.80	2.67	2.24	2.22
	[2.07]	[2.08]	[2.07]	[2.11]	[1.94]	[2.17]	[2.09]	[2.08]
Total	5.67	5.07	6.90	5.98	5.33	7.57	6.46	6.11
	[5.14]	[4.53]	[4.51]	[4.58]	[5.00]	[4.74]	[4.96]	[4.82]
Hard problems solved correctly								
Rounds 1–3 (Trial 1)	0.87	0.80	0.87	0.83	1.03	1.03	1.03	0.92
	[1.01]	[0.76]	[1.01]	[0.89]	[1.07]	[1.00]	[1.02]	[0.97]
Rounds 4–6 (Trial 2)	0.70	0.77	1.00	0.89	1.33	0.80	1.07	0.92
	[0.84]	[1.10]	[1.23]	[1.17]	[1.30]	[1.03]	[1.19]	[1.12]
Rounds 7–9 (Trial 3)	1.43	0.57	1.47	1.02	1.43	1.47	1.45	1.27
	[1.38]	[0.97]	[1.93]	[1.58]	[1.65]	[1.36]	[1.50]	[1.51]
Total	3.00	2.14	3.34	2.74	3.79	3.30	3.55	3.11
	[2.10]	[1.81]	[2.72]	[2.37]	[3.13]	[2.69]	[2.93]	[2.56]

Notes: *Time spent* is the main dependent variable and measures the number of seconds spent by a subject on solving the multiplication problems. *Time spent* is our operationalization of effort. *Problems solved correctly* measures the average number of multiplication problems solved correctly by a subject. *Problems solved correctly* is our operationalization of performance.

1,052.60 to 1,115.74 (or 6%) and further to 1,508.20 (or another 35%) when combined labels are used. Figure 3 provides a graphical depiction of effort.

To formally test whether the participants' efforts fall into the pattern predicted by Hypothesis 1, we use planned contrasts (Buckless and Ravenscroft 1990) given the directional prediction for more than two treatment groups. We use contrast weights of +3 for combined labels, −1 for positive-only labels, and −2 for no labels. The dependent variable is total time spent, that is, time spent over the three trials. Table 2, panel A, contains the test results, which support Hypothesis 1 regarding the prediction for effort ($F = 8.19$, $p < 0.01$, two-tailed).[13] Using other contrast weights, a nonparametric test or multilevel analysis leads to inferentially identical results.[14] The Wilcoxon-Mann-Whitney test confirms that effort is significantly higher in the two label conditions than in the no labels condition ($Z = 1.62$, $p = 0.05$, one-tailed).[15] However, pairwise comparisons of the combined labels, positive-only labels, and no labels treatments reveal that our results for effort in Hypothesis 1 are primarily driven by the difference of the combined labels treatment from the two other treatments. Using nonparametric pairwise tests, we find that effort in the combined labels treatment is significantly greater than effort in the positive-only treatment ($Z = 2.49$, $p < 0.01$, one-tailed) and in the no labels treatment ($Z = 2.52$, $p < 0.01$, one-tailed). However, we find no significant difference between the no labels and the positive-only treatment ($Z = 0.44$, $p = 0.33$, one-tailed). Thus, our prediction for effort in Hypothesis 1 is only partially supported.

Hypothesis 1 further predicts that performance follows the same pattern as effort if the effort-performance link is sufficiently strong. The descriptive results in Table 1, panel B, show an increase in performance of 5% from the no labels to the positive-only labels conditions (17.67 vs. 18.55) and an additional performance increase of 26% from the positive-only labels to the combined labels conditions (18.55 vs. 23.43). The smaller increase in performance (26%) than in effort (35%) from the positive-only labels to the combined labels conditions may be because solving harder problems is less sensitive to increases in effort and therefore puts a limit on the extent to which performance can improve with increased effort. This is evidenced by participants solving more of the easy (12.88 vs. 10.38) and medium problems (7.23 vs. 5.20) in the combined labels condition than in the positive-only labels condition, but about the same number of hard problems (3.32 vs. 2.97) in the combined labels condition versus the positive-only labels condition.

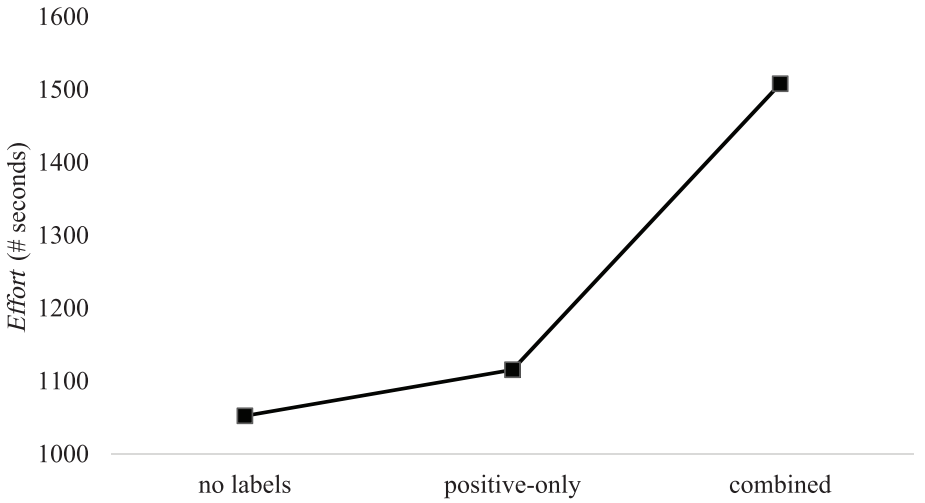
For the formal test of Hypothesis 1 for performance, we employ planned contrasts with the same weights as for our test for effort, and we use the sum of *Problems solved correctly* in the three trials as the dependent variable. Table 2, panel B, contains the results of this analysis, supporting the Hypothesis 1 prediction for performance ($F = 7.51$, $p < 0.01$, two-tailed). For

13. We report two-tailed test results, although we have a one-directional hypothesis for all *F*-tests since the *F*-distribution is nonsymmetric. Other test results for directional predictions are reported as one-tailed.

14. Our results for Hypothesis 1 (effort) are robust to the use of alternative contrast weights of +2, +1, and −3 ($F = 3.58$, $p = 0.06$, two-tailed) or the nonparametric Jonckheere-Terpstra test ($Z = 2.87$, $p < 0.01$, one-tailed). The results are also inferentially identical if we use effort in the first trial, that is, initial effort, as the dependent variable ($F = 6.10$, $p < 0.01$, two-tailed). Furthermore, the hypothesis holds regardless of the proportion of positively labeled ranks, that is, when the proportion of positively labeled ranks is either low ($F = 3.10$, $p = 0.08$, two-tailed) or high ($F = 7.78$, $p < 0.01$, two-tailed). Finally, we employ multilevel analysis, which accounts for correlated error terms resulting from having three trials of effort data nested within each subject and having effort data from six subjects nested within one group. Thus, in addition to a treatment variable for type of performance category label (with both proportion conditions being collapsed), the model includes random effects for trial and group. While we find significant treatment ($F = 6.68$, $p < 0.01$, two-tailed) and trial effects ($F = 14.29$, $p < 0.01$, two-tailed), the interaction term is insignificant ($F = 0.05$, $p = 0.99$, two-tailed). Most importantly, using the same contrast weights as for our main analysis, our results for the effect of effort predicted by Hypothesis 1 are inferentially identical to our main analysis ($t = 3.48$, $p < 0.01$, one-tailed).

15. We use a nonparametric test because the dependent variable is not normally distributed (Shapiro-Wilk test, $W = 0.91$, $p < 0.01$).

Figure 3 Mean effort in the no labels, positive-only labels, and combined labels conditions



Notes: Effort (# seconds) measures the number of seconds spent by a subject on solving the multiplication problems over all three trials of the experiment. In the no labels condition, no performance category labels were provided. In the positive-only labels conditions, the top ranks were labeled "good ranking position," and the bottom ranks were unlabeled. In the combined labels conditions, the top ranks were labeled "good ranking position," and the bottom ranks were labeled "poor ranking position."

performance, we again find that our results are robust to various other statistical test configurations.[16] However, pairwise comparisons of the combined labels, positive-only labels, and no labels treatments reveal that our Hypothesis 1 results for performance are primarily driven by the difference of the combined labels treatment from the two other treatments. Using nonparametric pairwise tests, we find that performance in the combined labels treatment is significantly greater than performance in the positive-only treatment ($Z = 2.49$, $p < 0.01$, one-tailed) and in the no labels treatment ($Z = 2.12$, $p = 0.02$, one-tailed). However, we find no significant difference between the no labels and positive-only treatments ($Z = 0.41$, $p = 0.34$, one-tailed). Thus, we conclude that Hypothesis 1 is only partially supported for performance. When developing Hypothesis 1, we further argued that we would find the same pattern for performance that we do for effort only if the effort-performance link is sufficiently strong. In line with this argument, if we use only the sum of medium and hard problems solved in the three trials as our dependent variable, that is,

16. Our results for Hypothesis 1 (performance) are robust to the use of an alternative contrast of +2, +1, and −3 ($F = 3.36$, $p = 0.07$, two-tailed) or the nonparametric Jonckheere-Terpstra test ($Z = 2.70$, $p < 0.01$, one-tailed). The results are inferentially identical if we use initial performance as the dependent variable ($F = 7.77$, $p < 0.01$, two-tailed). If we use only easy problems as our proxy for performance, the inferential results for initial performance ($F = 11.38$, $p < 0.01$, two-tailed) or total performance ($F = 12.47$, $p < 0.01$, two-tailed) still hold. The prediction of Hypothesis 1 for performance holds regardless of the proportion of positively labeled ranks, that is, when the proportion of positively labeled ranks is either low ($F = 4.30$, $p = 0.02$, two-tailed) or high ($F = 4.87$, $p = 0.03$, two-tailed). Finally, we employ multilevel analysis for the reasons described in footnote 14. While we find significant label ($F = 4.89$, $p < 0.01$, two-tailed) and trial effects ($F = 4.23$, $p = 0.02$, two-tailed), the interaction term is insignificant ($F = 0.60$, $p = 0.66$, two-tailed). Most importantly, using the same contrast weights as in our main analysis, our results for the effect for performance predicted by Hypothesis 1 are inferentially identical to our main analysis ($t = 2.99$, $p < 0.01$, one-tailed).

TABLE 2
Tests of hypotheses

Panel A: Contrast test for Hypothesis 1 (effort) (n = 150)

Dependent variable = Total *Time spent* in the three trials

Source of variation	df	MS	*F*-statistic	*p*-value (two-tailed)
Model contrast (contrast weights: combined = +3; positive-only = −1; no labels = −2)	1	5,665,156	8.19	<0.01

Panel B: Contrast test for Hypothesis 1 (performance) (n = 150)

Dependent variable = Total number of *Problems solved correctly* in the three trials

Source of variation	df	MS	*F*-statistic	*p*-value (two-tailed)
Model contrast (contrast weights: combined = +3; positive-only = −1; no labels = −2)	1	898	7.51	<0.01

Panel C: ANOVA model for Hypothesis 2 (effort) (n = 120)

Dependent variable = Total *Time spent* in the three trials

Source of variation	df	MS	*F*-statistic	*p*-value (two-tailed)
Label	1	4,620,903	6.69	0.01
Proportion	1	165,318	0.24	0.63
Label × *Proportion*	1	382,844	0.55	0.46
Residual	116	690,522		

Panel D: ANOVA model for Hypothesis 2 (performance) (n = 120)

Dependent variable = Total number of easy *Problems solved correctly* in the three trials

Source of variation	df	MS	*F*-statistic	*p*-value (two-tailed)
Label	1	188	7.45	<0.01
Proportion	1	2	0.08	0.77
Label × *Proportion*	1	3	0.11	0.74
Residual	116	25		

Panel E: Contrast test for Hypothesis 2 (effort) (n = 120)

Dependent variable = Total *Time spent* in the three trials

Source of variation	df	MS	*F*-statistic	*p*-value (two-tailed)
Model contrast (contrast weights: positive-only × low = −2; positive-only × high = −2; combined × low = +1; combined × high = +3)	1	5,145,472	7.45	<0.01

(The table is continued on the next page.)

TABLE 2 (continued)

Panel F: Contrast test for Hypothesis 2 (performance) (n = 120)

Dependent variable = Total number of easy *Problems solved correctly* in the three trials

Source of variation	df	MS	F-statistic	p-value (two-tailed)
Model contrast (contrast weights: positive-only × low = −2; positive-only × high = −2; combined × low = +1; combined × high = +3)	1	186	7.40	<0.01

we discard easy problems, Hypothesis 1 is not supported for performance (F = 1.66, p = 0.20, two-tailed) since performance becomes more dependent on ability.

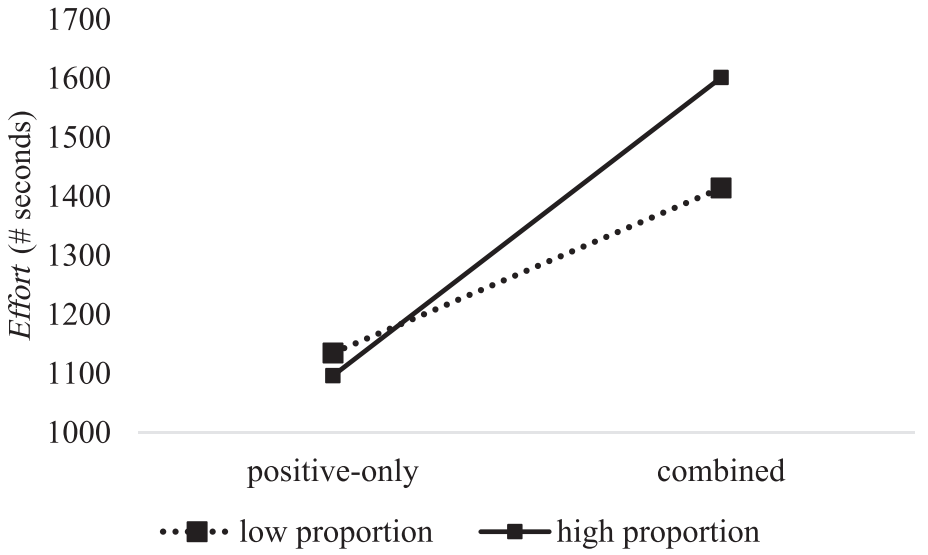
Hypothesis 2 predicts that the positive effect on effort of using combined, instead of positive-only, labels is stronger when the proportion of positively labeled ranks is larger. The descriptive statistics (Table 1, panel A), depicted in Figure 4, concord with this hypothesis: in the low-proportion treatments, the difference between effort in the combined labels condition (1,414.60) and that in the positive-only labels condition (1,135.10) is 279.50, whereas it is 505.43 in the high-proportion treatments (1,601.80 for combined labels; 1,096.37 for positive-only labels).

Table 2, panel C, reports the analysis of variance (ANOVA) results with type of performance category label (*Label*) and proportion of positively labeled ranks (*Proportion*) as the independent variables and total *Time spent* over the three trials as the dependent variable. We find that the interaction effect *Label* × *Proportion* is insignificant (F = 0.55, p = 0.46, two-tailed). However, because Hypothesis 2 predicts an ordinal interaction, we follow Buckless and Ravenscroft (1990) and formally test Hypothesis 2 using planned contrasts with weights consistent with our theory. We use contrast weights of −2 for the two treatments with positive-only labels. When the proportion of positively labeled ranks is increased, the top performance category becomes less prestigious, leading to lower motivation. This effect, however, is offset because, simultaneously, attaining the top performance category label appears to be more feasible; thus, individuals are willing to exert more effort. We use higher weights for the two combined labels treatments, consistent with our prediction in Hypothesis 1 that combined labels are more effective than positive-only labels. We employ a contrast weight of +3 for combined labels × high proportion (and +1 for combined labels × low proportion) because regulatory focus theory predicts that avoiding bottom ranks becomes a necessity when there are many positively labeled ranks. Table 2, panel E, contains the results, which support the Hypothesis 2 prediction for effort (F = 7.45, p < 0.01, two-tailed). The results are robust to further tests.[17]

Hypothesis 2 makes an analogous prediction for performance provided that the effort-performance link is sufficiently strong. More precisely, Hypothesis 2 predicts the difference in performance between combined labels and positive-only labels to be greater for a high proportion of positively labeled ranks than for a low proportion of such ranks. However, the descriptive results do not reveal the predicted pattern (Table 1, panel A). Although the difference between

17. The results are inferentially identical if we use effort in the first trial, that is, initial effort, as the dependent variable (F = 5.35, p = 0.02, two-tailed). We also employ multilevel analysis for the reasons described in footnote 14. We find significant treatment effects for labels (F = 9.32, p < 0.01, two-tailed) and trial effects (F = 13.02, p < 0.01, two-tailed), yet the treatment effects for proportion (F = 0.33, p = 0.56, two-tailed) and the interaction term are insignificant (F = 0.77, p = 0.38, two-tailed). Most importantly, using the same contrast weights as for our main analysis, our results for the effect for effort predicted by Hypothesis 2 are inferentially identical to our main analysis (t = 3.22, p < 0.01, one-tailed).

Figure 4 Mean effort in the two label conditions with a low versus a high proportion of positively labeled ranks



Notes: Effort (# seconds) measures the number of seconds spent by a subject on solving the multiplication problems over all three trials of the experiment. In the low-proportion conditions, the two top ranks of six ranks had positive labels. In the high-proportion conditions, the top four ranks of six ranks had positive labels. The remaining ranks are either unlabeled (positive-only labels) or negatively labeled (combined labels). Refer to Figure 3 for further variable definitions.

combined labels (22.83) and positive-only labels (17.60) is 5.23 in the low-proportion condition, we find a difference of only 4.53 in the high-proportion condition (24.03 for combined labels and 19.50 for positive-only labels). Thus, we find support for our prediction in Hypothesis 2 for effort but not for performance. Similar to our finding reported above that solving more difficult problems takes more time, we find that—on average for the four conditions—working one extra minute results in correctly solving an additional 2.3 easy problems but only 1.6 medium problems and 1.4 hard problems.[18] Participants who invest more time in solving the problems are more likely to start working on more difficult problems. However, solving these problems requires not only more effort but also more ability; that is, the effort-performance link is weakened. Although performance category labels have a positive effect on effort, there is no reason to believe that such labels also affect ability. Therefore, we operationalize performance with the number of *easy* problems solved correctly over the three trials when the effort-performance link is stronger. We find (Table 1, panel C) that, consistent with our theory, the increase in performance from positive-only labels to combined labels (10.40 vs. 12.60) is 21% in the low-proportion condition and 27% in the high-proportion condition (10.37 vs. 13.16). If we use this operationalization of performance (i.e., the sum of *easy* problems solved over the three trials) and apply planned contrasts with the same configuration that we used for the prediction of Hypothesis 2 for effort, we find support for the performance prediction ($F = 7.40$, $p < 0.01$, two-tailed), as shown in Table 2,

18. Wilcoxon signed rank sum tests confirm that the difference between easy (2.3) and medium (1.6) problems ($S = 2{,}034$, $p < 0.01$, two-tailed), the difference between easy (2.3) and hard (1.4) problems ($S = 1{,}655$, $p < 0.01$, two-tailed), and the difference between medium (1.6) and hard (1.4) problems are all significant ($S = 849$, $p < 0.01$, two-tailed).

panel F.[19] However, Table 2, panel D, also reports the ANOVA results showing an insignificant interaction effect ($F = 0.11$, $p = 0.74$, two-tailed).

We conclude that the positive effort and performance effects of combined labels versus positive-only labels are stronger if the number of positively labeled ranks is larger. With respect to performance, this outcome holds if the link between effort and performance is sufficiently strong.[20]

Additional analysis

This subsection provides further analyses and uses questions from the postexperimental questionnaire to further test the theory that underlies our hypotheses and to exclude fairness concerns as an alternative explanation.

Top, middle, and bottom performers

When developing Hypothesis 1, we argue that, irrespective of what performance category labels are used, these labels are particularly motivating for middle performers. While the status of top performers and bottom performers is less sensitive to slight changes in performance, middle performers should be particularly motivated to preserve or improve their status induced by performance category labels. It follows that middle performers should react more strongly to the presence of labels than top performers. As a direct test of our theory, we run the following regression model (which implicitly contains top performers as the benchmark of comparison):

$$\begin{aligned} Effort_{Trial\,2} = {} & \propto + \beta_1 BottomPerformer_{Trial\,1} + \beta_2 MiddlePerformer_{Trial\,1} + \beta_3 Label \\ & + \beta_4 BottomPerformer_{Trial\,1} \times Label + \beta_5 MiddlePerformer_{Trial\,1} \times Label, \end{aligned}$$

where

$Effort_{Trial\,2}$	Effort in the second trial (operationalized by *Time spent*);
$BottomPerformer_{Trial\,1}$	Dummy variable that equals one if the subject obtained rank 5 or 6 in the first trial, and zero otherwise;
$MiddlePerformer_{Trial\,1}$	Dummy variable that equals one if the subject obtained rank 3 or 4 in the first trial, and zero otherwise; and
Label	Dummy variable that equals one if (any type of) performance category labels are used, and zero otherwise.

We use only observations from the first two trials and not the third trial since some participants might be top performers in the first trial and bottom performers in the second trial. Thus, it is unclear whether these subjects should be classified as top, middle, or bottom performers, and predicting their effort in the third trial is ambiguous.

Table 3 shows that both interaction terms are significant. Hence, middle and bottom performers react more strongly to performance category labels than top performers do. Our finding for middle performers is consistent with our theory that performance category labels induce

19. In addition, we also employ multilevel analysis for the reasons described in footnote 14. We find significant treatment effects for labels ($F = 7.95$, $p < 0.01$, two-tailed), yet the treatment effects for proportion ($F = 0.00$, $p = 0.96$, two-tailed), trial effects ($F = 2.26$, $p = 0.13$, two-tailed) and the interaction term are insignificant ($F = 0.18$, $p = 0.67$, two-tailed). Most importantly, using the same contrast weights as for our main analysis, our results for the effect for performance (easy problems) predicted by Hypothesis 2 are inferentially identical to our main analysis ($t = 2.75$, $p < 0.01$, one-tailed).

20. Although Hypothesis 2 focuses on the interaction effect of the type of label category and the proportion of positively labeled ranks, the results also indicate that a larger proportion of positively labeled ranks leads to more effort and performance under combined labels. Since ranking systems with a higher proportion of positively labeled ranks can be considered more lenient, our results also suggest that more lenient forced rankings have a more positive influence than less lenient rankings. However, we are careful in interpreting and generalizing this effect since a simple main effect test for the combined labels condition provides no significant effect (untabulated).

TABLE 3
Regression analysis of the label effect for top, middle, and bottom performers (n = 150)
$Effort_{Trial\ 2} = \propto + \beta_1 BottomPerformer_{Trial\ 1} + \beta_2 MiddlePerformer_{Trial\ 1} + \beta_3 Label + \beta_4 BottomPerformer_{Trial\ 1} \times Label + \beta_5 MiddlePerformer_{Trial\ 1} \times Label$

Variable	Parameter estimate	(*t*-value)
$\propto$	623.60***	(7.38)
$BottomPerformer_{Trial\ 1}$	−549.40***	(−4.60)
$MiddlePerformer_{Trial\ 1}$	−363.80***	(−3.04)
Label	−106.73	(−1.13)
$BottomPerformer_{Trial\ 1} \times Label$	303.68**	(2.27)
$MiddlePerformer_{Trial\ 1} \times Label$	310.03**	(2.32)
Adj. R^2	0.2065	
F-value	8.76***	

Notes: $Effort_{Trial\ 2}$ is the effort in the second trial (operationalized by time spent). $BottomPerformer_{Trial\ 1}$ is the dummy variable that equals one if the subject obtained rank 5 or 6 in the first trial, and zero otherwise. $MiddlePerformer_{Trial\ 1}$ is the dummy variable that equals one if the subject obtained rank 3 or 4 in the first trial, and zero otherwise. *Label* is the dummy variable that equals one if performance category labels are used, and zero otherwise. ** and *** denote statistical significance at the 5% and 1% levels, respectively (two-tailed).

greater competition for status for middle performers than for top performers. The positive regression coefficient implies that performance category labels result in greater effort for middle performers.[21] The results show that the $BottomPerformer_{Trial\ 1} \times Label$ interaction is also significant. Although bottom performers—similar to top performers—are less likely to receive a rank with a different status, the performance category matters. This outcome is in line with our prediction that social comparisons are intensified, and employees become more competitive when label information is available. As individuals usually engage in upward instead of downward comparisons, this intensification is more pronounced for bottom performers than for top performers.[22]

Social comparison concerns

In developing our hypotheses, we also argue that social comparisons increase in the presence of performance category labels and that this increase is even more pronounced when combined labels are used, rather than positive-only labels. We capture relative performance concerns via three postexperimental questions, as suggested by prior research (Hannan et al. 2013, 2019;

21. For a deeper understanding of the behavior of middle performers, we replicated the analysis and investigated effort in the third trial. To avoid mixed experience in the previous two trials, we only considered subjects whose status (in terms of being a top, middle, or bottom performer) had not changed throughout trials 1 and 2 (98 observations). The results (untabulated) show that the interaction of being a middle performer and the presence of a label is significantly positive (p = 0.07, two-tailed) as in the main analysis, while the interaction of being a bottom performer and the presence of the label lacks significance (p = 0.31, two-tailed). This underlines the importance of the behavior of the middle performers.

22. A follow-up Kruskal-Wallis test shows no significant differences in the efforts of top performers when no labels, positive-only labels or combined labels are used (p = 0.39, two-tailed). Furthermore, we find that the coefficients of the two interaction terms ($BottomPerformer_{Trial\ 1} \times Label$ and $MiddlePerformer_{Trial\ 1} \times Label$) are not significantly different. Although we argue that performance category labels are particularly informative (and thus motivating) for middle performers, the shame associated with being labeled as a bottom performer comes into play as a second force for motivating bottom performers. This finding potentially explains why the coefficients are not significantly different.

Tafkov 2013).[23] Table 4, panel A, shows the participants' responses to the respective questions and our aggregated measure *Social Comparison Factor* (eigenvalue = 1.90; explained variance = 63.4%). We conduct a planned contrast test (untabulated) using this factor score as the dependent variable and the same weights used in Hypothesis 1 to test whether social comparisons follow the same pattern as effort and performance. As predicted by our theory, we find that social comparisons are lowest when no labels are used, that they increase in the presence of positive-only labels and that they are greatest when combined labels are employed ($F = 4.05$, $p = 0.05$, two-tailed). Therefore, social comparisons follow the same pattern predicted by Hypothesis 1.

With respect to Hypothesis 2, we test whether social comparisons caused by using combined labels, instead of positive-only labels, are more pronounced when the proportion of positively labeled ranks is high instead of low. This test (untabulated) lends further support to our theory that underlies Hypothesis 2. We use the same contrast weights that we used in testing Hypothesis 2 and find that social comparisons follow the predicted pattern ($F = 3.39$, $p = 0.07$, two-tailed). This outcome supports our reasoning that adding combined labels (versus adding positive-only labels) increases social comparison concerns to a greater extent when the proportion of positively labeled ranks is larger.

Shame focus

We further argue that adding negative labels to the bottom ranks establishes a prevention focus that becomes more important when the proportion of positively labeled ranks is larger. We contend that the shame associated with a low ranking and employees' motivation will both increase with a higher proportion of positively labeled ranks. Therefore, we asked the participants the extent to which they thought about the shame associated with ranks 5 and 6 (1 = *very little* and 11 = *very much*). Table 4, panel B, contains the descriptive results. Consistent with our theory, we find that (results untabulated) using combined labels instead of positive-only labels causes the participants to think significantly more about shame when the proportion of positively labeled ranks is high ($Z = 1.81$, $p = 0.04$, one-tailed) than when it is low ($Z = 0.40$, $p = 0.34$, one-tailed).[24]

Other prevention focus effects

Regulatory focus theory also predicts that a prevention focus affects *how* employees pursue their work. When a prevention focus is established, employees have a "concern for avoiding mistakes, because errors are seen as costly and ominous. Such a concern causes prevention-focused individuals to work slowly and to be overly diligent" (Johnson et al. 2010, 231). In contrast, when only a promotion focus is present, employees instead "adopt an eagerness strategy that emphasizes speed" (Johnson et al. 2010, 231). Therefore, we calculate two measures, accuracy and diligence, to test the predictions of regulatory focus theory. We define accuracy as the number of problems solved correctly over the number of problems that the participants attempted to solve. We define diligence as the number of seconds that the participants spend thinking about a problem before entering a solution. We find that both accuracy and diligence are higher in the combined treatments, which add a prevention focus, than in the positive-only treatments. More precisely, accuracy (diligence) is, on average, 49.78% (29.90 seconds) in the positive-only treatments and

23. On an 11-point Likert scale, the subjects were asked: (i) the extent to which they were nervous about their relative performance, (ii) whether thinking about how their performance compared to the performance of other participants interfered with their ability to concentrate on the problems, and (iii) how often they thought about their ranking relative to other participants.

24. If we use the same contrasts that we used to test Hypothesis 2, we find the same pattern; that is, the increase in the perceived shame of bottom-ranked positions if combined labels are used is greater for a high proportion than for a low proportion of positively labeled ranks ($F = 2.96$, $p = 0.09$, two-tailed).

TABLE 4
Responses to postexperimental questions (mean, [SD])

Panel A: Social comparison questions and factor (n = 150)

		Low proportion		High proportion	
	No labels	Positive-only	Combined	Positive-only	Combined
Were you nervous or concerned about how well you were performing (total number of problems correctly solved) relative to other participants in the experiment?	2.87 [2.10]	3.37 [2.62]	3.47 [2.56]	3.90 [3.21]	4.67 [3.19]
Did thinking about how your performance (total number of problems correctly solved) compared to other participants interfere with your ability to concentrate on the problems?	3.27 [2.55]	3.93 [2.78]	4.00 [2.70]	3.73 [2.82]	4.50 [2.62]
How often did you think about how your performance (total number of problems correctly solved) ranked relative to other participants in the experiment?	5.67 [2.91]	5.40 [2.92]	5.70 [2.55]	4.73 [2.64]	6.40 [2.86]
Social Comparison Factor (calculated)	−0.22 [0.92]	−0.06 [1.03]	0.01 [0.87]	−0.10 [1.10]	0.37 [1.02]

Panel B: Other questions (n = 150)

		Low proportion		High proportion	
	No labels	Positive-only	Combined	Positive-only	Combined
To what extent did you think about the shame associated with ranks 5 and 6?	3.70 [3.23]	3.77 [3.32]	4.23 [3.69]	3.60 [3.35]	5.07 [3.72]
How fair do you perceive the rank-based evaluation to be?	6.70 [3.02]	7.40 [2.37]	6.27 [3.16]	7.10 [2.75]	6.13 [2.81]

Notes: Panels A and B contain questions and answers from the postexperimental questionnaire. Answers were provided on an 11-point Likert scale with higher numbers representing greater importance/shame/fairness. Panel A also contains a *Social Comparison Factor* calculated by applying principal component analysis to the answers to the three questions (explained variance 63.42%; eigenvalue = 1.90).

55.30% (34.14 seconds) in the combined treatments (descriptive statistics untabulated). A Wilcoxon-Mann-Whitney test (untabulated) confirms that accuracy and diligence are significantly higher when combined labels are used than when positive-only labels are used (Z = 1.68, p = 0.05, one-tailed for accuracy and Z = 1.56, p = 0.06, one-tailed for diligence). Therefore, in line with regulatory focus theory, a prevention focus motivates employees to avoid mistakes and to think more thoroughly before entering a solution. However, this outcome also affects the subjects' trade-off between approaching as many problems as possible and solving them correctly, which could serve as an additional explanation for why we find only limited support for Hypothesis 2 regarding performance; if the subjects spend too much time ensuring that they entered the correct solution, performance increases less than effort.

Fairness concerns

Finally, we consider fairness as an alternative explanation for Hypothesis 1 because it could explain why effort and performance increase when negative labels are added to positive labels in the combined labels condition. Prior research argues that contracts perceived as fairer result in greater effort (Hannan et al. 2005). If employees perceive that it is unfair that bottom ranks are *not* "punished" by a negative label in the positive-only labels treatment, the labeling is perceived as unfair and is therefore less effective. The postexperimental questionnaire contains a question about the fairness of the information system (1 = *very unfair* and 11 = *very fair*). Table 4, panel B, shows the results. We find that the information system is perceived as fairer when positive-only labels are used than when combined labels are used (low proportion: 7.40 vs. 6.27; high proportion: 7.10 vs. 6.13). A Wilcoxon-Mann-Whitney test (untabulated) confirms that, over both proportion conditions, the difference between positive-only and combined labels is significant ($Z = 1.88$, $p = 0.06$, two-tailed), but the difference between positive-only and no labels is not ($Z = 0.86$, $p = 0.39$, two-tailed). While these results show that fairness does not explain our results, the lower perceived fairness of combined labels could serve as a potential explanation for why many firms in practice refrain from using combined labels.

5. Conclusion

In this study, we investigate whether combining fine ranking (RPI) with coarse ranking via performance category labels increases effort and performance. We examine two important facets of performance category labels that firms must determine. First, we examine positive-only and combined labels. Second, we analyze whether a higher proportion of positively labeled ranks affects effort and performance differently under combined labels than under positive-only labels.

We find that adding performance category labels increases effort and performance if top and bottom ranks are labeled but not if only top ranks are labeled. The performance effects depend on a sufficiently strong effort performance link. We attribute our findings to social comparisons, status concerns, and regulatory focus theory. While our results would imply that firms should prefer combined over positive-only labels, positive-only labels seem more common in practice (Charness et al. 2014; Dohmen 2012). Our additional analysis reveals a potential explanation. We find that participants perceive combined labels as less fair. Thus, employers may respond by offering combined labels contracts less frequently.

Furthermore, we find that the positive effect of using combined labels instead of positive-only labels is more pronounced when the proportion of positively labeled ranks is larger. This occurs because being a member of the low-status group is a greater burden when this group is smaller. Again, our results for performance hold only if the effort-performance link is sufficiently strong.

From a theory perspective, our study expands our understanding of the effectiveness of (unrewarded) performance category labels when employees already have detailed ranking information. Whereas Charness et al. (2014) document in a flat-wage setting negative performance effects of combined performance category labels that are provided in addition to RPI, we find positive effects in a setting in which a piece-rate incentive scheme motivates individuals to care about their performance. Our findings also have important implications for management accountants involved in the design of performance information systems. First, our results demonstrate the opportunity costs of firms that use positive-only, instead of combined, labels, which they must trade off with employees' preferences for positive-only labels. Second, our findings illustrate that both the type of category label used and the proportion of the categories are important decisions for firms because they affect effort and performance.

Future studies might further explore this field of research. Although our study demonstrates that the presence and type of performance category labels matter for effort and performance, future research could investigate the use of subjective, instead of objective, performance information. A potential limitation of our study is that it requires homogeneity

not only *within* groups but also *across* groups. Otherwise, an employee with good absolute performance might fall into a poor performance category simply because competition in one group is tougher than in other groups. Future research might further investigate this issue and examine the effectiveness of performance category labels added to RPI if employees are less homogenous or performance category labels are awarded based on absolute, instead of relative, performance. While we find that using combined labels becomes even more beneficial when the proportion of positively labeled ranks increases, this relationship may not be linear. In other words, if employees are almost certain to receive a positively labeled rank, complacency might set in, implying a potential boundary condition of the reported results. Future research might therefore further examine the effectiveness of different levels for the proportion of positively labeled ranks.

References

Baiman, S. 1982. Agency research in managerial accounting: A survey. *Journal of Accounting Literature* 1 (1): 154–213.

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

Berger, L., K. Klassen, T. Libby, and A. Webb. 2013. Complacency and giving up across repeated tournaments: Evidence from the field. *Journal of Management Accounting Research* 25 (1): 143–67.

Besley, T., and M. Ghatak. 2008. Status incentives. *American Economic Review* 98 (2): 206–11.

Bhattacharya, H., and S. Dugar. 2012. Status incentives and performance. *Managerial and Decision Economics* 33 (7–8): 549–63.

Bhattacharya, H., and S. Dugar. 2013. Contests for ranks: Experimental evidence. *Southern Economic Journal* 79 (3): 621–38.

Brockner, J., and E. T. Higgins. 2001. Regulatory focus theory: Implications for the study of emotions at work. *Organizational Behavior and Human Decision Processes* 86 (1): 35–66.

Buckless, F. A., and S. P. Ravenscroft. 1990. Contrast coding: A refinement of ANOVA in behavioral analysis. *The Accounting Review* 65 (4): 933–45.

Carbery, G. 2012. Ryanair denies fuel claim after Maydays. *Irish Times*, August 17, https://www.irishtimes.com/news/ryanair-denies-fuel-claim-after-maydays-1.537994, retrieved July 10, 2020.

Caruth, D. L., and G. D. Handlogten. 2001. *Managing compensation (and understanding it too)*. Westport, CT: Quorum Books.

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

Dohmen, F. 2012. Brutal psychological terror: Taking on employee intimidation at T-Mobile USA. *Spiegel International*, November 22, https://www.spiegel.de/international/business/union-campaign-takes-on-t-mobile-usa-working-conditions-a-868525.html, retrieved July 10, 2020.

Eddington, K. M., F. Dolcos, R. Cabeza, K. R. R. Krishnan, T. J Strauman. 2007. Neural correlates of promotion and prevention goal activation: An fMRI study using an idiographic approach. *Journal of Cognitive Neuroscience* 19 (7): 1152–62.

Festinger, L. 1954. A theory of social comparison processes. *Human Relations* 7 (2): 117–40.

Fischbacher, U. 2007. Z-tree: Zurich toolbox for ready-made economic experiments. *Experimental Economics* 10 (2): 171–78.

Förster, J., E. T. Higgins, and C. L. Idson. 1998. Approach and avoidance strength as a function of regulatory focus: Revisiting the "goal looms larger" effect. *Journal of Personality and Social Psychology* 75 (5): 1115–31.

Frederickson, J. R. 1992. Relative performance information: The effects of common uncertainty and contract type on agent effort. *The Accounting Review* 67 (4): 647–69.

Hannan, R., V. Hoffman, and D. Moser. 2005. Bonus versus penalty: Does contract frame affect employee effort? In *Experimental Business Research*, edited by A. Rapoport and R. Zwick, 151–69. Dordrecht, The Netherlands: Springer.

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

Hannan, R. L., G. P. McPhee, A. H. Newman, and I. D. Tafkov. 2013. The effect of relative performance information on performance and effort allocation in a multi-task environment. *The Accounting Review* 88 (2): 553–75.

Hannan, R. L., G. P. McPhee, A. H. Newman, I. D. Tafkov, and S. J. Kachelmeier. 2019. The informativeness of relative performance information and its effect on effort allocation in a multitask environment. *Contemporary Accounting Research* 36 (3): 1607–33.

Hazels, B., and C. M. Sasse. 2008. Forced ranking: A review. *SAM Advanced Management Journal* 73 (2): 35–39.

Hecht, G., I. D. Tafkov, and K. L. Towry. 2012. Performance spillover in a multitask environment. *Contemporary Accounting Research* 29 (2): 563–89.

Higgins, E. T. 1997. Beyond pleasure and pain. *American Psychologist* 52 (12): 1280–300.

Higgins, E. T. 1998. Promotion and prevention: Regulatory focus as a motivational principle. In *Advances in Experimental Social Psychology*, edited by M. P. Zanna, 1–46. New York: Academic Press.

Higgins, E. T. 2002. How self-regulation creates distinct values: The case of promotion and prevention decision making. *Journal of Consumer Psychology* 12 (3): 177 –91.

Huberman, B. A., C. H. Loch, and A. Önçüler. 2004. Status as a valued resource. *Social Psychology Quarterly* 67 (1): 103–14.

Johnson, R. E., C.-H. Chang, and L.-Q. Yang. 2010. Commitment and motivation at work: The relevance of employee identity and regulatory focus. *Academy of Management Review* 35 (2): 226–45.

Knauer, T., F. Sommer, and A. Wöhrmann. 2017. Tournament winner proportion and its effect on effort: An investigation of the underlying psychological mechanisms. *European Accounting Review* 26 (4): 681–702.

Kramer, S., V. S. Maas, and M. v. Rinsum. 2016. Relative performance information, rank ordering and employee performance: A research note. *Management Accounting Research* 33: 16–24.

Lazarus, R. 1991. *Emotion and adaption*. New York: Oxford University Press.

Lipman, V. 2012. The pros and cons of forced rankings: A manager's perspective. *Forbes*, July 19, http://www.forbes.com/sites/victorlipman/2012/07/19/the-pros-and-cons-of-forced-rankings-a-managers-perspective/, retrieved July 10, 2020.

Loftus, S., and L. Tanlu. 2017. Because of "because": Examining the use of causal language in relative performance feedback. *The Accounting Review* 93 (2): 277–97.

Moldovanu, B., A. Sela, and X. Shi. 2007. Contests for status. *Journal of Political Economy* 115 (2): 338–63.

Murphy, K. R., and J. N. Cleveland. 1995. *Understanding performance appraisal*. Thousand Oaks, CA: Sage Publications.

Newman, A. H., and I. D. Tafkov. 2014. Relative performance information in tournaments with different prize structures. *Accounting, Organizations and Society* 39 (5): 348–61.

Pennington, G. L., and N. J. Roese. 2003. Regulatory focus and temporal distance. *Journal of Experimental Social Psychology* 39 (6): 563–76.

Peterson, S. J., and F. Luthans. 2006. The impact of financial and nonfinancial incentives on business-unit outcomes over time. *Journal of Applied Psychology* 91 (1): 156–65.

Ridgeway, C. L., and H. A. Walker. 1995. Status structures. In *Sociological Perspectives on Social Psychology*, edited by K. Cook, G. Fine and J. House, 281–310. New York: Allyn and Bacon.

Seher, D. 2012. Ryanair soll seine Piloten zum Kerosin Sparen nötigen ("Ryanair said to force their pilots to save kerosene"). *NRZ*, August 16, https://www.nrz.de/reise/ryanair-soll-seine-piloten-zum-kerosin-sparen-noetigen-id6989962.html?page=2, retrieved July 10, 2020.

Shah, J. H., and E. T. Higgins. 2001. Regulatory concerns and appraisal efficiency: The general impact of promotion and prevention. *Journal of Personality and Social Psychology* 80 (5): 693–705.

Smith, R. 2000. Assimilative and contrastive emotional reactions to upward and downward social comparisons. In *Handbook of Social Comparison: Theory and Research*, edited by J. Suls and L. Weheel, 173–200. New York: Kluwer Academic/Plenum Publishers.

Sprinkle, G. B. 2000. The effect of incentive contracts on learning and performance. *The Accounting Review* 75 (3): 299–326.

Tafkov, I. D. 2013. Private and public relative performance information under different compensation contracts. *The Accounting Review* 88 (1): 327–50.

Tesser, A. 1988. Toward a self-evaluation maintenance model of social behavior. In *Advances in Experimental Social Psychology*, Vol. 21, edited by L. Berkowitz, 181–228. New York: Academic Press.